THE UNIVERSITY *of* EDINBURGH

Know Yourself Better In and Through Peer Disagreement

James Joseph Kraft

Candidate for Doctor of Philosophy
University of Edinburgh, 2023

ABSTRACT: I aim to answer the following question in this dissertation: How far can one trust claims to self-knowledge based on privileged access in epistemic peer disagreements where those claims are the focus of the disagreement? The answer to this question is contained in the following Assessment Framework:

> **Assessment Framework:**
> In a peer disagreement where the privileged self-knowledge claim of one disputant is crucially consequential for the disagreement, trust the claim prima facie only if there are no or little significant signs of "judgmental awareness" and/or of observational evidence that implies the claim is questionable; and adjust credence in the privileged self-knowledge claim according to the following scale of such significance: No signs = highest credence; little signs = high credence; significant signs = low credence; highest signs = lowest credence.

This Assessment Framework is derived both from six crucial components (numbered in parenthesis below) of peer disagreements about privileged self-knowledge claims discussed throughout the chapters. I begin by arguing, based on empirical research, that both the observational-interpretive method for knowing oneself (component 1) and the privileged access method (2) are needed, and they work together as integrated with mindfulness (3) to form more reliable privileged self-knowledge claims. We show how scholars of peer disagreements make mistakes by not seeing how privileged self-knowledge claims are fragile (4). The remaining crucial components are the Prima Facie Norm (5), which says we should accept prima facie a privileged claim, and the Indirect Scrutability Norm (6) which says that the privileged claim can be scrutinized indirectly by the observational-interpretive method. The culminating sixth chapter derives the Assessment Framework from the six key factors, on one hand, and tests the Assessment Framework against four case studies given, on the other hand. The measure of the success of this framework is how well it naturally accounts for ordinary lived experiences. With the Assessment Framework the person with the privileged self-knowledge claim and the person critical of the privileged claim complete the expanded Delphic maxim Know Yourself (γνῶθι σεαυτόν) better <u>in</u> peer disagreements. <u>Through</u> the process of specifying how to better know yourself <u>in</u> peer disagreements you learn that humans need both ways of knowing yourself, that you need the feedback of others to help you make sure your mental states are what you think they are, that you can know yourself better through mindfulness, and that peer disagreements about privileged self-knowledge claims have value in that they are one of the best ways to deeply Know Yourself. So, the Delphic maxim used also by Socrates and Plato is extended to come to the following: Know Yourself (γνῶθι σεαυτόν) better <u>in</u> and <u>through</u> peer disagreements about privileged self-knowledge claims with mindfulness both in privileged and observational-interpretive access integrated.

Signed Declaration:
I declare that this thesis has been composed solely by myself, James Joseph Kraft, and that it has not been submitted, in whole or in part, in any previous application for a degree or a professional qualification. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own. There are no included publications.

LAY SUMMARY: This thesis is about a special kind of self-knowledge, the kind only one person can have about herself. For example, only I can know for sure whether I meant to insult someone or did it by accident. Others can infer, but only I can know my true intentions in an immediate way. Let's call this special self-knowledge a "privileged self-knowledge claim." This thesis explores how well I can know my own privileged self-knowledge claim when someone equivalently knowledgeable and skilled at evaluating the situation disagrees with me.

Here's the answer broken down into a guideline which I call the Assessment Framework: Trust the privileged self-knowledge claim of a disputant at face value only if there's no significant signs of rash judgments and of observational evidence that suggests the claim may be wrong. For example, suppose a friend of mine insists that I intentionally insulted a mutual friend. I respond, "I did not intend to insult him." My friend says this privileged self-knowledge claim is false. She believes this based on her observation that I rashly and automatically denied that the insult was intended. And my friend insists she has observational evidence that implies this privileged claim is false, like seeing me in the past clearly intending an insult even though I denied intent. She says these two sources of evidence support her view that, while I might think I didn't intend the insult, I deep down did. If these two pieces of evidence are true, the Assessment Framework recommends that I should have less trust that my privileged self-knowledge claim is true.

While this Assessment Framework is derived and tested in the final chapter, the previous chapters provide the understandings necessary for its derivation. Chapters One through Three establish three things. First, using the help of mindfulness studies and studies about self-observation, it can be shown that there is such a thing as privileged self-knowledge. Second, the same studies assert that humans also need to know themselves through observation and not just through privilege access. Third, that privilege self-knowledge claims can be mistaken given the many psychological, cultural, and physical barriers to privileged self-knowledge. Fourth, mindfulness can help us determine when a person is less likely to have a privileged self-knowledge claim that is true, for example when the person makes rash and automated decisions. The next two chapters describe two principles needed for the Assessment Framework to work: the idea that privileged self-knowledge claims are often wrong; and the idea that privileged self-knowledge claims can be scrutinized through observational evidence that suggests the person's claim may be wrong.

An implication of the Assessment Framework used in peer disagreement is that peer disagreements about privileged self-knowledge can help us know ourselves better. A good peer in a disagreement can observe my actions and speech and tell me when my judgments are rash and when my behaviors suggest that my claim may be false.

Table of Contents

INTRODUCTION

I aim to answer the following question in this dissertation: How far can one trust claims to self-knowledge based on privileged access in epistemic peer disagreements where those claims are the focus of the disagreement? Before describing the answer argued for, let's get some definitions out in the open:

Self-knowledge:	Knowledge of one's own mental states and attitudes, e.g., one's beliefs, jealousy, or intentions.
Privileged access:	A way of knowing one's own mental states and attitudes that is direct, without observation, without inference, and uniquely had by the one who has the mental state or attitude.
Epistemic peer disagreement:	A disagreement with a person who is roughly as likely to get the answer right because that person has comparable knowledge about the details of the issue, and is comparable capable of evaluating those details.
Privileged self-knowledge claim:	A sincere claim to self-knowledge based on privileged access to one's mental state.
Target disagreement:	A disagreement with an epistemic peer over one disputant's self-knowledge claim based on privilege access.

The answer to this question I argue for is contained in the Assessment Framework immediately below for assessing how far to trust a particular privileged self-knowledge claim in a particular disagreement.

Assessment Framework:
In a peer disagreement where the privileged self-knowledge claim of one disputant is crucially consequential for the disagreement, trust the claim prima facie only if there are no or little significant signs of "judgmental awareness" and/or of observational evidence that implies the claim is questionable; and adjust credence in the privileged self-knowledge claim according to the following scale of such significance: No signs = highest credence; little signs = high credence; significant signs = low credence; highest signs = lowest credence.

This framework can only be understood adequately in Chapter Six where it is derived both from six crucial components of target disagreements discussed throughout the chapters and from four crucial interrelations of those components (Both are listed at the end of this introduction.). The measure of the success of this framework is how well it naturally accounts for ordinary lived experiences like the one's portrayed in the four case studies given farther below which exemplify typical target disagreements.

As a preview of what can only be fully understood later, we can say one of the most important things about this Assessment Framework. It recognizes the following normative starting point and normative ending point for trust in claims to self-knowledge based on privileged access in peer disagreements: Trust such claims prima facie since they normally reflect the reality of one's mental states when sincerely stated, and don't trust them when there are signs of "judgmental awareness" and observational-interpretive evidence known to each that

implies the claim is questionable. The scare quotes "judgmental awareness" require explanation. "Judgement" here doesn't refer to specific judgments that we have to make in our daily lives, e.g., that the checker is competent, that I will be safe walking down this street at night, etc. Rather, "judgmental awareness" refers to the particular quality of the awareness in which everyday judgments are formed, that is, an awareness in which one is rash to judge what one is aware of.

While the Assessment Framework is developed in Part Three of this text, the six key components used to derive the Assessment Framework are developed in Part One and Part Two. The three chapters of Part One provide the foundational understanding of four of these six key components of target disagreements: the fragile nature of privileged self-knowledge claims, mindfulness, and the two methods of knowing oneself that will be the basis for all the insights about target disagreements that follow. The two ways of knowing oneself are through observational-interpretive access and privileged access. For an example of the observational-interpretive method, I observe my stomach growling and I infer that I am hungry. For an example of the privileged access method of knowing oneself, consider the person who says, after deliberating about all the possibilities, that she is not jealous. In Part One I prove, with empirical studies, that people really do attain self-knowledge through both the privileged access method and the observational-interpretive method, that mindfulness is strongly and positively correlated with privileged self-knowledge, and that the two methods for knowing oneself are integrated with the help of mindfulness. From that foundation I conclude, in Part Two, that the scholarly literature on peer disagreement doesn't adequately recognize both the fragile nature of privileged self-knowledge claims and the integration of the two methods of knowing oneself that can help mitigate the said fragility thus attaining more reliable assessments of the comparative epistemic status of disputants in target disagreements. You will see that Part Two also offers two remedies for these oversights in the literature, remedies that integrate the two methods for knowing oneself with mindfulness. The two remedies presented in the Second Part are the fifth and sixth components of target disagreements: the Prima Facie Norm and the Indirect Scrutability Norm. These six key components of target disagreements described in the five chapters of Part One and Part Two are the building blocks used in Chapter Six of Part Three for the Assessment Framework.

With the Assessment Framework derived and verified to be useful in Part Three, both the person with the privileged self-knowledge claim (the claimant) and the person critical of the privileged claim (the claim-critic) complete the expanded Delphic maxim Know Yourself (γνῶθι σεαυτόν) better in peer disagreements, the claimant because she knows (γνῶ) better the epistemic boundaries of her own privileged claim in responding to the challenging concerns of her peer, and the claim-critic because she knows (γνῶ) better the epistemic boundaries of her criticism in responding to the challenging concerns of her peer.

"γνῶθι σεαυτόν" has the verbal form "γνῶθι" in the second person, singular, aorist, active, imperative conjugation and the word "σεαυτόν" in the accusative, singular, and second person. It is an urgent and personal request/command addressed to one person, that is, the second person singular. That person is you, and me right now as the one individual reading the two words over two thousand years after they were etched in stone. The aorist in the verb implies that the request/command calls for you to do something in the present that is an appropriate response to something that took place in the past. The request is not for something to keep going, as in "Keep signing" (Palmer, 2023). It is a request/command to start something now that hasn't been done before and so isn't something that has already started.

While it is customary to translate "σεαυτόν" as Thyself, I will not do so. The word "thyself" is too formal, and the Delphic maxim is a personal and informal request/command. The word "yourself" preserves the informal aspect. The translation "thyself" avoids a little ambiguity in the English word "yourself," which sometimes is used incorrectly to refer to a group of people, as in "You should all train yourself [which should be yourselves] to sell the product faster;" it seems most people know that "yourself" can only refer to one person; so, for the most part the translation "Know Yourself" preserves the personal and individual appeal/command of "γνῶθι σεαυτόν." Also, to preserve the personal and individual nature of the Delphic maxim, I will, when specifically trying to express its personal appeal/command, use the second person singular pronoun where traditional academic writing encourages the use of "one" or "we."

While the derivation and successful application of this Assessment Framework is a major goal of this dissertation, we also learn something deeper about self-knowledge through peer disagreements which is implied by the Assessment Framework and all the work done to derive it. I take there to be a distinction between

Knowing Yourself in versus through target disagreements. "Knowing Yourself in" specifies a particular context that particular self-knowledge can be gained within. Indeed, most of this dissertation will be talking about how, in the use of the Assessment Framework, the claimant and the claim-critic know better the epistemic boundaries of their claim and claim-criticism respectively for the particular disagreement. But, for this Assessment Framework to work, an essential interrelatedness must be assumed between what one knows about oneself through privileged access and what one knows about oneself through observational-interpretive access, as we will see in Chapter Six. Thus, through the process of specifying how to better know yourself in peer disagreements you learn this interrelatedness leads to the conclusion that humans need both ways of knowing yourself, a conclusion that has wide implications explored briefly in the conclusion of this dissertation. Secondly, you learn through target disagreements that you need the feedback of others to help you make sure your mental states are what you think they are. Thirdly, through the process of deriving this Assessment Framework you also learn that you can know yourself better through mindfulness. Finally, you come to see that target disagreements aren't necessarily the sort of thing you drudgingly engage in. Rather, they have value in that they are one of the best ways to deeply Know Yourself. So, in the conclusion of this dissertation the Delphic maxim used also by Socrates and Plato is extended to come to the following deeper understanding of the epistemic human condition as regards self-knowledge: Know Yourself (γνῶθι σεαυτόν) better through target disagreements with mindfulness both in privileged and observational-interpretive access integrated.

Since the measure of the success of this Assessment Framework is how well it accounts for lived experiences, which we will later represent in four case studies that reflect typical and complex peer disagreements over self-knowledge claims based on privileged access, it would be good here to disclose those four case studies at the outset:

> **JEALOUS: "I know I am not jealous!"**
> Kamal talks to his longtime therapist about difficulties he has been having with a person at work whom he hates because this person is mean. He sees his therapist as an impartial epistemic peer on interpersonal issues. He describes the many ways that this person is mean in different cases. The therapist tells him that she thinks he, rather, hates the person because he's jealous. Kamal's irritation is clear when he says, "I know I am not jealous!" His therapist says,
>
> > "You describe that your colleague has achievements that you value and have not accomplished. We have talked about some instances in the past when you conceded jealousy when initially you insisted you weren't. Sometimes you don't pay attention to how your emotions influence your thoughts. You say you are not a detail person; you don't like to pay attention to distractions like clocks ticking, birds chirping, and smells of things. You say when driving you don't care to notice things if they aren't matters of safety. I don't think you would be able to observe how your emotions influence your thoughts in this stressful situation. I can see them because I am, unlike you, impartial in this case. Furthermore, your reactivity to the idea that you are jealous was impulsive without hearing me out or even taking the view seriously; that may indicate you don't want to hear the evidence, that it is a defense mechanism triggered to help you think good things about yourself."

Kamal remembers the incidents she is talking about where he finally conceded jealously, but he thinks this case is significantly different. Sometimes his therapist is wrong. While he thinks the past instances of similar mistakes are inconclusive, he finds compelling the point that he often doesn't see the extent to which his emotions influence his view of others, given that he isn't a detail person. And it is telling how he was so reactive. He has less confidence in his view.

LOVE: "I know I love you!"

Blake says to his wife Anna, "I love you." Anna says, "No you don't; many of the things that you do and say indicate that you don't love me; though I know that you sincerely think you do love me!" At this point Blake sees his wife as an epistemic peer. He is eager to hear her out because he knows her to be very sincere and perceptive. Anna describes how he yells at her often about her not being able to get along with her supervisors at work, degrades her for not being able to drive and insisting that he taxi her and their son wherever she wants to go, complains about her obsessions, and berates her in public. While Blake acknowledges that he has done these things and that these things often indicate the absence of love, he believes there are extenuating circumstances. He says the acrimonious behaviors result from his inability to cope with the negative consequences—her not being able to keep a job to support a child, not being able to drive, obsessions, and excessive worries—that result from her three disabilities officially diagnosed by a psychologist: Attention Deficit Disorder of the Inattentive Type, Generalized Anxiety, and Delusional Disorder of the Non-Bizarre Type. He also thinks she isn't doing enough to address her difficulties, and to help others who have difficulties with her three disabilities. To his therapist he describes his acrimonious behavior, and he also describes how she always quickly reacts to his acrimonious behaviors in an automated and angry way without taking his concern for extenuating circumstances seriously, claiming he isn't a good person, just selfish. His therapist explains how such a knee-jerk reaction often is a sign of a defense mechanism which makes it less likely to correctly understand another's mental state by preventing consideration of alternative views or information which may be true and may soften one's criticism. At the end, he comes to be steadfast in reasoning both that she just isn't adequately able to see his legitimate extenuating circumstances, and that her interpretation of him as not a good person and selfish is influenced by her documented Delusional Disorder.

RACIST: "I know I am not racist."

You are a white faculty member at a large public university, and you are charged with other colleagues to decide a tenure review for another colleague who is African American. You are terrified of making a racist decision, but at the same time you are equally dedicated to upholding standards of teaching and scholarship that you believe are good for all races and for the success of your university. At the preliminary tenure review committee meeting, you express your concerns about this case: "He is close to fulfilling the standards of scholarship that we need to uphold the quality of our PhD program, but he misses the mark." Later that day at a bar you frequent, your friend and tenure committee colleague, Samantha, wants to talk to you about what you said. She helped you skillfully get through many interpersonal and academic problems, and so you consider her an epistemic peer on this issue. She says, "I agree with you that his scholarship isn't the greatest. But, it is a little better than you estimate. I am wondering if you have considered that you may be biased in the estimate of his scholarship?" You say, "If you are implying that I am racist, you are wrong. I know I am not racist. I have many friends that are African American, and I have never discriminated due to race." Samantha says, "I am not saying you are KKK racist, or Jim Crow racist. But, based on my observation, I think you might have a little of what Chris Rock at the Oscars called 'Sorority Racism' where you have black friends and you love the idea that they have equal opportunities, but you don't want them in your inner group." She points out how you avoid African Americans at conference receptions, you don't come to the bar when you know an African American faculty person is likely to be there. You listen carefully for every instance of avoidance she cites, and you find extenuating circumstances. You point out how there is one African American colleague at conferences you always avoid because that person doesn't work on the issue you write on, and at conferences you mostly only want to talk to people writing on your topic. And you avoid the African American colleague at the bar because she always wants to talk about sports. You then point out how Samantha must not have seen how you like talking to particular African Americans at conferences and this bar you frequent. You can't see anything in your speech or behavior that indicates you don't want these African American colleagues in your "sorority." What you say makes good sense to Samantha. In fact, she then recalls one time when you were extensively and naturally talking shop with a different African American colleague at a conference. She conciliates, "I am sorry I thought you were a sorority racist, but I still think you underestimate the person's scholarship."

INTENT: "I know that I didn't intend to kill my husband!"

Chris is a juror in a murder trial that hinges on whether the defendant, Sam, intended to kill her husband. He listened to her testimony:

> I know that I didn't intend to kill my husband! Such intent goes against everything that I am. At a time when he was always at the office and there were a spate of daytime home invasions in our neighborhood, I was taking the gun out of my purse in order to clean it after target practice. Suddenly, I see and hear someone reaching around my shoulder trying to grab my hand. [*Now crying and pausing to recover enough composure to talk*] He must not have seen the snub nose revolver in my hand. When he pulled my hand towards him, the gun went off. I didn't know at that point that it was my husband. I feared for my life that I was being attacked. As he laid dying, I saw it was my husband, and he said, "I just wanted to dance with you like we always did when we were doing well."

There is a good amount of evidence contradicting her testimony: They were in the middle of a terrible custody battle for their children; there is no gun residue on her husband's hand, which is expected when near a gun firing; the manager at the range testified that he complimented her about how well she was shooting the target, to which she replied, "If only it were my husband;" two of her colleagues recall her saying a day before the incident in the break room after crying about the difficulties with her husband, "I should just kill him;" the blood splatter, position of his body when he fell, and the entry point of the bullet were all highly inconsistent with her testimony. And there were a few scratch marks on her hand and legs as if a fight happened. An expert crime scene investigator testified that he must have been shot from at least six feet away. Given all this evidence, Chris believes she probably murdered her husband, but he also thinks there is reasonable doubt. Perhaps her shirt was between his hand and the gun such that he didn't get gun residue. And he just thinks there is no way that she could have faked the extremely deeply felt and sincere testimony she gave. Then, a well-respected forensic psychologist who specializes in defendant testimony made in extremely stressful situations takes the stand and describes many critically acclaimed studies she and her colleagues have published that statistically prove something about people who, on one hand, are in extreme psychologically stressful situations like Sam, and, on the other, do something in that situation that goes against their self-concept. Most of the time they sincerely believe they didn't intend to do it even when they actually did intend to do it at the time. The psychologist also points to tell-tale signs that Sam is repressing the memories that threaten her understanding of herself. She forcefully denies having said, "If only it were my husband," and "I should just kill him" even when there appears to be no bias in those who testified to the remarks. The forensic psychologist says based on the evidence they have found and the relevant

psychological testing they have had her do, she is suffering from Brief Psychotic Disorder with a stressor (DSM-5 298.8). Because of all this additional evidence Chris now believes Sam intended to murder her husband.

These case studies above are diverse. One is about a privileged self-knowledge claim about jealousy; a second about love; a third about racism; and a fourth about intent to kill. In each of these a privileged self-knowledge claim is the core focus of the disagreement. In each the privileged self-knowledge claim disputed matters vitally.

Notice also that each case study focuses on one of the four most important perspectives of a peer disagreement about privileged self-knowledge. The first, JEALOUS, focuses on the perspective of the person holding the privileged self-knowledge claim when it is clear that she should be conciliatory, that is, reducing the credence level of the privileged self-knowledge claim; and from here on we will refer to this person as the "claimant." The second case study, LOVE, also presents the perspective of the claimant when it is clear that she should be steadfast, that is, not reduce the level of credence of the privileged self-knowledge claim. Correspondingly, the third case study, RACIST, presents the perspective of the person critical of the claim when it is clear that she should be conciliatory about the credence level of her critique of the privileged claim; and from here on we will call this person the claim-critic. The fourth case study, INTENT, also focuses on the perspective of the claim-critic, only this time it is clear that she should be steadfast.

Before describing each chapter in more detail, an overview of the chapters can be given in terms of how they support the Assessment Framework and its derivation from the six key components and four key interrelations. The first five chapters describe the six key components of target disagreements, while the culminating sixth chapter derives the Assessment Framework from the six key factors and their four key interrelations, on one hand, and tests the Assessment Framework against the four case studies given above, on the other hand. The following are the key components and their interrelations which will help us derive the Assessment Framework in Chapter Six.

Six key components influencing peer disagreements about privileged self-knowledge claims

1) The fragility of privileged and observational-interpretive access due to the many psychological, cultural, and physical barriers.
2) The observational-interpretive access method for gaining self-knowledge of mental states in peer disagreement.
3) The privileged access method for gaining self-knowledge of mental states in peer disagreement.
4) Mindfulness.
5) The Prima Facie Norm.
6) The Indirect Scrutability Norm.

Four key interrelations among key components:

Interrelation #1: The two norms are interrelated because the two methods of knowing oneself are interrelated.
Interrelation #2: Mindfulness improves both the methods of knowing yourself.
Interrelation #3: Each disputant can observe signs of higher and lower mindfulness.
Interrelation #4: Each disputant can have four different types of observational-interpretive evidence.

PART ONE

Observational-Interpretive Access and Privileged Access Integrated with Mindfulness

In Part One I argue that empirical studies prove there are two methods needed for gaining self-knowledge, one through observation/interpretation of oneself and another through privileged access. I also argue that empirical research shows that these two methods are integrated with mindfulness, integrated in the sense that mindfulness allows one to better differentiate between them, to appreciate each for what it uniquely does, and to use them together more effectively.

These two methods and their integration with mindfulness are the main focus of the three chapters of this first part. In the process of describing observational-interpretive models and their critique of privileged access, Chapter One argues for the need for the observational-interpretive method and differentiates the latter from the privileged access method. Chapter Two argues each method is better differentiated and better appreciated for what it uniquely does with the help of mindfulness, and it does this in the process of arguing that mindfulness uncovers the limits of observational-interpretive models. Chapter Three argues that when the two methods work together more effectively with mindfulness, we can see that adequate mindfulness is needed for privileged self-knowledge claims and that the latter are limited. In doing what I have described just above, Part One builds the foundation of the Assessment Framework of self-knowledge claims based on privileged access in peer disagreements. The Assessment Framework developed in Chapter Six will be shown to be the epitome of the integration with mindfulness of the two methods for gaining self-knowledge since it effectively uses together both methods with mindfulness for the process of assessing the epistemic status of each disputants' perspective in a peer disagreement.

Part One contributes crucially to the derivation of the Assessment Framework in Chapter Six which is a major goal of this dissertation. As discussed, the Assessment Framework is derived from six key components of peer disagreements about privileged self-knowledge claims (see the list "Six key components" above). In

developing an understanding of the first four of the six key components, this Part One provides the backbone for everything we talk about in subsequent chapters. Those four components once again: The fragility of privileged self-knowledge claims, the observational-interpretive access method for gaining self-knowledge, the privileged access method, and mindfulness. The groundwork for the Prima Facie Norm discussed in Chapter Five is developed here along with a deep understanding of how mindfulness fosters self-knowledge.

Chapter One

Observational/Interpretive Access Models and Their Critique of Privileged Access

Chapter One argues that we need the observational-interpretive method for attaining self-knowledge in the process of describing observational-interpretive models of self-knowledge and their critique of privileged access. With the help of empirical research on self-observation, we prove that the observational-interpretive method is needed for attaining self-knowledge. We first discuss what a privileged self-knowledge claim is. We then describe the extreme position that there is never any privileged access, and one of the best representative works of this extreme is that of Peter Carruthers. While there are others who pursue this observational-interpretive model (e.g., Gopnik 1993, Cassam, 2014), Carruthers represents well the outlines of this approach. We then argue that empirical studies on self-observation prove that observational-interpretive models are right to assert the need for the observational-interpretive method for attaining self-knowledge, even though they don't support the observational-interpretive models' view that observational-interpretive access is the only way to gain self-knowledge. There is nothing in the research that says observation is the only way to attain self-knowledge, though we truly do need the observational-interpretive method for attaining self-knowledge. To find out why observational-interpretive models think observation is the only way to self-knowledge, we show how Carruthers, one of the most well-known proponents of the observational-interpretive model, defends his never-any-privileged-access generalization both against Richard Moran's deliberative agency view and against the view, described in his 2007 article ("The Illusion of Conscious Will"), that system-two thinking involves privileged access. Carruthers describes four criteria that self-knowledge and the process of attaining it would have to have if that self-knowledge is privileged. In this process we clearly differentiate the observational-interpretive method from the privileged access method. In the next two chapters, we will argue against this model that says there is only the observational-interpretive method for knowing one's metal states, and we will do this by describing self-knowledge that meets these four criteria Carruthers presents for privileged self-knowledge.

Carruthers' work is representative of the outlines of observational-interpretive models. A key component of observational-interpretive models of self-knowledge is that one doesn't know directly one's mental states, that the observations of things in one's internal (like thoughts in inner speech) and external environment cue one to believe one has a particular mental state. Another key component of observational-interpretive models is interpretation. For Krista Lawlor one has internal promptings whose cause one then interprets to give one self-knowledge (Lawlor, 2009). Internal promptings can be sensations, imaged sentences, or visual images. Lawlor gives an example of a mother who, looking over her baby in a crib, observes her thought, "Have another." That thought observed has to be interpreted. Quassim Cassam also thinks that we make inferences from the internal promptings that we observe (Cassam, 2014, 143). He calls his view inferentialism. Daryl Bem's work is probably the best example of an observational-interpretive model, as he thinks one attributes attitudinal mental states to oneself that are implied by one's observable behaviors (Bem, 1972). For example, I observe a growl coming from my stomach and I come to believe that I am hungry. Here, I don't access the mental state immediately from my own agency; it is cued indirectly through the observation of the growling stomach. Carruthers' views differ from many other observational-interpretive accounts with his view that substantial self-knowledge is interpreted unconsciously. Though other observational and interpretive accounts may not think substantial self-knowledge is determined unconsciously, they all think one doesn't have immediate, direct, and conscious access to one's attitudinal mental states. There is no direct access because there are always intermediary components like inference, conscious interpretation, unconscious interpretation, etc.

In this chapter we are focused on instances of sincere claims to self-knowledge based on privileged access. People usually assume that others have exclusive access to their own mental states. Of course, if we suspect that someone is lying, we don't trust that person, not because we don't think that person has privileged access to her mental states, but, rather, because we suspect she is intentionally not telling others what she indeed has privileged access to. In this study we are interested only in such claims that are sincere.

One extreme view: There never is privileged access

Carruthers' argument in his Interpretive Sensory Access (ISA) view is that nobody ever has privileged access to their own substantial mental states because people always lack two components necessary for such access, namely, non-sensory and non-interpretive access. Here's the argument in a nutshell:

The main argument of ISA:

- If access to substantial mental states is privileged, then such access is not sensory and not interpretive.
- But, such access is always sensory and interpretive.
- So, access to substantial mental states is <u>never</u> privileged.

We can express this argument which is, to a certain extent, supported by the results of the empirical studies on self-observation we are about to present: Access is sensory and interpretive because the information coming in from perception, inner speech, concepts used, personal reasoning, etc., is processed by unconscious systems that observe and interpret the information forming substantial attitudes, decisions, and judgments without the person's awareness. In what follows we will quickly describe Carruthers' specific argument, and then show how relevant empirical studies on self-observation support to a certain extent Carruthers' argument, even if they don't support his generalization represented in "The main argument of ISA" above with the word "never."

We can quickly describe what is behind Carruthers' generalization that there is <u>never</u> any privileged access to one's own substantial mental states by describing the three points that Carruthers' says his major book is about (Carruthers, 2011, starting at 1-2). First, the access one has to one's own substantial mental states is essentially the same as the access one has to the mental states of others. The only way that one can know the mental states of others is to engage sensory observation of their actions, speech, demeanor, etc. And the results of the observations of the other are interpreted in order to come to an understanding of what the person's substantial mental states are. Carruthers says essentially the same observational-interpretive and interpretive processes are used when we come to know our own substantial mental states. One faculty is used to understand both the mental states of others and of oneself, and he calls this the mindreading faculty. And it makes sense for Carruthers from an evolutionary perspective that there is one mindreading faculty for both, since the most important thing for the evolving hominid is to understand the substantial mental states of others so that one can

figure out whether to trust them and since evolution doesn't select for new processes if existing ones are adequate (Carruthers, 2011, 67-68).

Secondly, we can see the justification for Carruthers' "never" by recognizing his view that, just as the mindreading of others can only involve sensory observation, so too the mindreading of ourselves can only as well. All the sensory inputs—coming from perception, the five senses, concepts, inner speech rehearsing ideas, proprioception, interoception, etc.—are fed to the mindreading faculty through what he calls the global broadcast system, following the work of Baars (Baars, 1988) (Carruthers, 2011, 47-48). The mindreading faculty gets unconscious sensory inputs as well (Carruthers, 2011, 52). One's thoughts and inner speech are fed to the mindreading faculty as inputs. Substantial decisions and judgments are not in the sensual information, though some non-substantial judgments in the perceptual context are transparent and able to be accessed in a privileged way, like judgments about the valence, the texture, and the shape of the things we experience. So, Carruthers allows for perceptually bound judgements and decisions to be known in a privileged way, though, again, substantial judgments and decisions can't be known in a privileged way. Sensory bound judgments and decisions aren't substantial because they don't do the heavy work for us, interpreting how we should respond to our perceptions. Inner speech in and of itself doesn't make substantial decisions even though it gives us conscious content and can be factored in when the mindreading faculty determines what one believes; it is just streams of different ideas fed to the mindreading system which will there be evaluated and serve as objects that the mindreading system has access to through the global broadcast system. That information is observed by specialized (Carruthers, 2011, 47) conceptual and affective systems that consume and draw inferences from that information. Visual imagery is also fed to the global broadcast system. Internal speech is conscious, reflective, and, so, system-two reasoning; but it gets processed with unconscious system-one processes, which are quick, non-reflective, and don't use working memory. One then consciously goes along with the substantial decisions and judgments resulting from the different systems accessing the same inputs. He describes this broadcast system as like specialists in a room receiving information on a common blackboard that all the systems have access to in order to accomplish their specific functions. The specialists can only communicate with one another by writing messages on the board, which are visible to all (Carruthers, 2011, 48). There is no

higher authority organizing the specialists (Carruthers, 2011, 49). Because mindreading evolved for outward-looking social purposes, the only inputs that the mindreading process receives are perceptual and imagistic (Carruthers, 2011, 69). In fact, whenever we entertain a current thought, it is grounded in sensory awareness of our circumstances, behavior, and sensory items held in working memory (Carruthers, 2011, 4). We can have transparent and privileged access only to the content, modality, and perceptually embedded judgments of our sensory context (Carruthers, 2011, 76-78), not to substantial judgments and decisions that require interpretation. We cannot know our attitudes transparently by looking outwardly (Carruthers, 2011, 118). The mind is opaque and hidden; hence the title of his book, *The Opacity of Mind*.

A third major conclusion is that the essentially unconscious system-one mindreading process is always engaged in generating consequential and substantial propositional attitudes, and that means that any access is interpretive. Two things show us that our access to our substantial mental states is interpretive: The fact that when we just look at the things themselves, we don't see substantial decisions and judgments in them; they show up as dogs and cars, not heaps of molecules. Second, the fact that people often confabulate especially when there is ambiguity in a situation or when a situation presents unsettling negative information about the self. There are many different reasons why people confabulate. People often unconsciously interpret the situation in a way favorable to their bias, self-image, or desires with false memories or false details that make them look better, without even knowing that the confabulations are taking place. The profile of confabulation results from experiments where sensory information usually unambiguously indicating a decision or judgment is manipulated to produce ambiguity, or experiments where there is no sensory information that helps the subject know what the unconscious decision is. These profiles of confabulation match exactly what you would expect with an ISA model of self-knowledge: manipulate the sensory cues for a decision and you get confabulations because the subject has the wrong sensory cue for a decision (Carrruthers, 2011, 340); when you take away sensory information allowing subjects to know what the decisions are, subjects have to rely on other means of finding a decision or judgment like bias and independent theoretical inferences, and this leads to confabulations (Carruthers, 2011, 341) (See Nisbettt and Wilson, 1977). This is known as choice blindness. We might think that we consciously interpret the things around us, and that we come to know our substantial

mental states through the decisions and judgments we consciously make. But Carruthers brings up many different examples of how this certainly doesn't happen. The point of his presentation of the many ways in which people are mistaken in their self-knowledge attributions due to false unconscious interpretations isn't to prove the skeptical view that we never have self-knowledge, but rather to prove the ISA view that all access to substantial mental states is interpreted from sensory inputs (Carruthers, 2011, 68). The argument for ISA is an empirical one, since if there were non-inferential access to our beliefs, empirical research would not show so much confabulation in the empirical research on confabulation, like it does. Carruthers thinks we sometimes have self-knowledge given the ubiquitous interpretive-sensory access. Indeed, one of the best empirical researchers on self-knowledge, Timothy Wilson, cites research supporting the crucial importance better observation of one's own sensory information has for producing reliable self-attributions of mental states (Wilson and Dunn, 2004).

Empirical self-observation studies partially support the never-have-privileged-access view

Whether or not you agree with Carruthers' granular view of mindreading, the general outline of his ISA is to a certain extent supported by many scientific studies on self-observation, even though this empirical research doesn't support Carruthers' generalization that we never have privileged access. Here we will prove with the results of empirical studies that sensory and interpretive access is widespread, and thus can't be ignored when understanding the access we have to our mental states. Certainly, empirical studies say that self-attributions of mental states are often based on sensory observation.

In all the examples that we are about to describe, people are mistaken in their self-knowledge claims based on sensory self-observation because something about the two following components of self-observation is wrong, ambiguous, limited, or manipulated: the sensory cues and the interpreted significance of the sensory cues. A cue is a configuration of things observed in one's sensory environment set up in advance to trigger a particular interpretation of that configuration.

A cue …

is a configuration of things observed in one's sensory environment set up in advance to trigger a particular interpretation of that configuration.

Here is an example of a cue: a long piece of strong wood at one end going through to the end of a much smaller piece of metal at a 90º angle such that the entire thing looks like the letter "T" with a flat circular surface about one inch on one end of the metal. That configuration of things in one's environment, the wood in relation to the metal, in our culture is set up to trigger the interpretation "It's a hammer." In what follows we will see examples of self-knowledge claims that are mistaken because they are based on the observation of sensory cues where something about the components of self-observation is wrong, ambiguous, limited, or manipulated. From these examples of people in diverse and widespread cases making self-knowledge claims based on interpretive-sensory access going wrong, the empirical research concludes, like Carruthers, that we <u>often</u> and in diverse situations make self-knowledge claims based on interpretive-sensory access. The only difference with Carruthers' conclusion, compared to the results of empirical studies, is that he generalizes with the "never," as we will see.

Let's start with a 1966 study by Stuart Valins where men were asked to view pictures of minimally dressed women when they thought falsely they were at the same time listening to their own heart rate amplified in their headphones (Valins, 1966). When men falsely thought they were hearing their own heart rate rapidly increase as they observed a particular photo, they reported the mental state of liking the woman in the photo. What they in fact were listening to was a prerecorded heart rate. This is evidence for thinking at least some self-attributions of mental states depend on, and are mediated by, observed sensory information that gets interpreted unconsciously. Wilson says, "Behavior change often precedes changes in attitudes and feelings" (Wilson 2002, 89). Indeed, in the Valins study even the observation of alleged behavior changes produces attitude change. Sometimes claims to self-knowledge based on observing oneself get it right; and sometimes claims to self-knowledge based on observing oneself get it wrong. And, as Wilson and others claim, studying how this process gets it wrong can help us understand how this process gets it right.

The fact that sensory-based cues influence mental states and the knowledge of them is evidence that we know our minds often indirectly through the sensory cues found in the sensory context (Carruthers 2011, 339). One study starts out with people given headphones, ostensibly to evaluate the headphones themselves, with a recording playing of a person giving an argument on an issue; those randomly induced to nod were more

likely to agree with the argument presented (Wells and Petty, 2008). Nodding one's head is a sensory cue customary in many cultures indicating one approves of what is being discussed; it is a cue both for others observing the nodding and for oneself actually doing it. The person's observation of a common cue, here randomly manipulated, produces the mental state of a positive judgment regardless of what a person believed about the argument before.

Research after Valins has confirmed the importance of sensory observation for determining mental states and the knowledge of them. One study from 2013 by Steven Makkar and Jessica Grisham demonstrates that people with social anxiety assess their level of anxiety based on the observation of sensory information about the self in the sensory context such as heart rate, sweating, or blushing (Makkar and Grisham 2013). Those who falsely observe the heart rate go up in a given speech task, have more dysphoria and ruminations, and more self-focused attention. The observation of increased heart rate, sweating, etc., serves as cues that automatically and unconsciously generate an interpretation whether the cue organically occurred or was manipulated. It seems reasonable to think people who observe the cue of their own heart rate increasing, and not a manipulated cue, can actually know indirectly through non-privileged sensory observation that their anxiety level is rising so long as higher levels of anxiety are reliably associated with increased heart rate. This is clear evidence that the observation model of self-knowledge of Carruthers often is accurate. We often do have unconscious, system-one, sensory, and interpretive access to our mental states.

A similar study confirms the importance of self-observation for determining what one feels or thinks (Gray et al. 2007). Subjects were presented with pictures of neutral faces while their brain functions were observed in an fMRI. Perceived emotional intensity increased in those subjects with false feedback of increased heart rate. These studies can only mean that people rely on sensory observations of themselves, whether false or real, for determining what they feel or think. This process also is of the system-one type, quick and unconscious.

Gauging one's propositional attitude immediately based on cues visible to all is often a good short-hand way of coming to know one's most important propositional attitudes. We can't live without system-one immediate and unconscious thinking. When I see a tiger in the wild, I don't proceed to ponder the question of

whether I am fearful by consciously and deliberately answering the question. Rather, I yield to the interpretation of my mental state cued up to trigger when such an experience happens.

Consider another early groundbreaking study on these matters and the subsequent confirming research. The work of Daryl Bem supports the view that the observations people make of their own behavior influence the mental states that they have. In 1970 Bem described a study he did which measured the attitudes of college students about whether or not students should have control of their curriculum (Bem 1970). They reported their attitudes before and after they actually wrote an essay arguing that students should not have control of their curriculum. People whose attitudes initially were for such control changed their attitudes after writing the essay against control. Not even recognizing that they had changed their attitudes, students confabulated that they always had the belief that students shouldn't have such control. Here again an automated interpretation of the sensory cue can get in the way of understanding and remembering one's earlier mental states. The individual concludes from the observation of writing an essay against control that she had this attitude all along, which she didn't, inferring from her behavior that she must be against the proposal, since people writing such essays with little compensation must have the attitude of the position argued for. Bem found that people who didn't have any rewards were more likely to harmonize their attitudes according to their observation of their behavior. Subjects thought they were using their conscious rationality to make the decision when they were in fact engaged in system-one processes. One might say that those in the process of writing the essay perhaps found good reasons not to let students have control over their curriculum, and they just forgot their earlier attitudes; no need to think they are just following a sensory-interpretive cue. Even if this were the case, though, this wouldn't explain away the fact that subjects without a reward were more likely to harmonize their attitudes according to the automated understanding of their behavior.

Bem's fundamental view that observation of behavior produces attitude changes is confirmed by subsequent research. For example, James Laird in 1974 concludes that emotions (anger, happiness, including liking, disliking) change depending on the facial expressions people observe in themselves. Simply smiling can induce the mental state of feeling happy (Laird, 2007). In 1981 researchers Chaiken and Baldwin conclude that the remembrance of times when one was environmentally conscious affected what one believes about their

environmental attitudes regarding whether they think of themselves as environmentalists (Chaiken and Baldwin, 1981). Changes in the observation salience of memories influences attitudes. In 2006 Tiffany Ito and colleagues find that just inducing a person to smile when presented with a picture of someone of another race, in this case African American, makes the person report less bias against people of that other race (Ito et al. 2006). Guadagno and colleagues in 2010 find that people harmonize their attitudes towards radicalization according to the level of radical behavior they observe themselves engaging in (Guadagno et al, 2010). In 2010, Clayton Critcher's and Thomas Gilovich's research concludes that people determine their attitudes of approval or boredom depending on the observation of their mind wandering (Critcher and Gilovich 2010). Just the observation alone of oneself saying "I am excited" when attempting a stressful activity like public speaking changes the attitude one has towards the event (Brooks, 2014).

Here is another now classic example supporting the view that attributions of mental states are often mediated by observational-interpretive cues, this time, one that is culturally influenced. Nisbett and Wilson devise an experiment where subjects are asked to judge which panty hose have the most quality (Nisbett and Wilson 1977). Subjects preferred the items on the right even though they actually were indistinguishable. The best explanation for the propensity to error, in these examples but also in the examples to follow, is that there is an intermediary unconscious process of interpretation generally going on in the act of self-attributing one's own substantial and most important mental states.

The sensory cues which Bem, Valins, Carruthers and Wilson have proven to be often used for coming to understand our occurrent mental states themselves are the activation points—triggers—for decisions unconsciously made deep within the psyche regarding how to process things in the sensory context. Observing oneself nod one's head or accelerate the heart rate are triggers for the unconscious decisions "I like this argument" or "I like this picture" respectively. Why and how particular combinations of sensory data come to activate or trigger pre-established decisions is worked out unconsciously in the deep recesses of the psyche accessible only indirectly through sensory feedback.

Here is what every single one of these empirical studies on self-observation and self-knowledge taken together confirm with a high degree of scientific confidence:

What research on self-observation says:

> For a large range of the substantial mental states that I come to know about myself, I come to know them in the same way that others come to know my substantial mental states, that is, by sensory observation of cues and information that trigger automated and unconscious interpretations about my mental states.

Carruthers in his ISA says the exact same thing with one huge exception, he says that sensory/interpretive access is the only way one can know oneself. Thus, the empirical studies on self-observation corroborate much of what Carruthers says in his ISA, even though it doesn't support his "never" generalization. Though you might not agree with Carruthers' details about what is going on unconsciously with a global broadcast of sensory images relayed equally to different sub-processes and subsequently interpreted unconsciously, he does give us a generally good understanding of how the mind is often opaque with regard to substantial mental states.

System-one thinking is extremely important, vital actually. Though we need this automated and unconscious process for coming to know our mental states, and though it often results in the right mental state that reflects who we are, it can go wrong, as we have seen in the studies presented. The link between the cue and the mental state comes apart when something about the two following components of self-observation is wrong, ambiguous, limited, or manipulated: the sensory cues, and the interpreted significance of the sensory cues.

Carruthers' generalization in his critique of Moran-type thinking

Opposing Carruthers, Richard Moran thinks we normatively do have privileged access to our mental states, like a belief or a decision, resulting from our deliberative agency. Referring to first-person privileged access to judgments without observation and evidence, he says:

> Suffice it to say that they are taken to have a good prima facie claim to truth which may be overruled only in special cases. The important point is that these are taken to be genuine judgments, expressive of knowledge, which are made without reliance on "external" observation. (Moran, 2001, 10)

So, for Moran a privileged self-knowledge claim (recall that this means: sincere claims to self-knowledge based on privileged access) is rightly prima facie granted as a good claim to truth, and that grant ought to be overruled "only in special cases" (Moran, 2001, 10). Later, I too conclude privileged self-knowledge claims ought to be granted prima facie status; only I argue the prima facie status is much more fragile and conditional than Moran acknowledges in the quote above. Carruthers argues that Moran is simply wrong about this. In his book *The*

Opacity of the Mind, he discusses many of the ways of thinking about self-knowledge that Moran and others use in their defense of first-person authority. He describes conditions that would have to be fulfilled in order for there really to be a rationally deliberative and substantial decision, judgment, or commitment of an agent with privileged access. We shall list them here and then discuss why he thinks these conditions can't in principle be met. Later, we describe why Moran thinks these conditions are met, and we will see how mindfulness supports Moran's prima facie view. Carruthers' stated conditions for a truly, rationally deliberative, and privileged decision, judgment, or commitment of an agent:

Privileged Access: rationally deliberative, substantial decision, judgment, commitment of an agent must …

1) **Not be causally generated by an unconscious mental state (Carruthers, 2011, 96).**
2) **Not be generated by an interpretation (Carruthers, 2011, the entire book is about this).**
3) **Not made on the basis of observation and inference to a mental state (Carruthers, 2011, the entire book is about this).**
4) **Show transparently one's mental state in the process of settling an issue (Carruthers, 2011, 83).**

We can think of these four criteria as necessary conditions of privileged access.

Carruthers argues that these conditions can't in principle be met, given the right understanding of the mindreading faculty. The consequential decisions and judgments are made unconsciously downstream of any conscious rational reason or motive falsely thought to be what completely motivates a decision, judgment, or commitment. Rational conscious deliberation does have a place in Carruthers' ISA view influencing the resulting belief coming from the mindreading system. It just doesn't have the last say as to what one believes about one's own mind or the mind of others. For example, suppose a person rationally thinks about all the factors regarding the ethics of abortion, and comes to say the following in inner speech, "Abortion is unjust." That statement made in conscious deliberation is broadcasted to the mindreading system and taken into consideration by this unconscious system along with all the information of interoception and proprioception, and then one commits to the belief in post-conscious deliberation, or doesn't depending on its druthers. One's belief on the matter isn't completely settled until the final, decisive, and unconscious deliberation is made after any conscious deliberation. Even the deepest consciously thought-out decisions don't really settle the matter, since they are followed by the truly consequential and unconscious decisions in the mindreading system one is

in principle not aware of. Substantial decisions, judgments, and commitments aren't determined by the agents; rather, they are determined by unconscious processes of the mindreading system.

There is no transparency in the origin of decisions, judgments, and commitments. Moran thinks that we find out what our decisions and judgments are in the process of settling an issue. This is called transparency (Moran, 2001, 63). Carruthers thinks there is transparency in perceptually embedded judgments, but there is no transparency when it comes to substantial decisions and judgments. Were transparency the conduit for knowing one's judgments and decisions that would mean the judgments are produced without the interpretive influence of the unconscious mindreading system. You would know your judgments in real time as one settles an issue without unconscious influence. But you don't know your judgments and decisions transparently in settling an issue; the research on confabulation, ulterior unconscious motives, and self-observation proves this isn't the case (Carruthers, 2011, first chapter, 340-341), hence the "interpretive" component of Carruthers' Interpretive Sensory Access (ISA).

The fact that humans are prone to confabulation and all sorts of epistemic-debilitating automated processes discussed in the second section leads us to the understanding that we humans don't have the ability to decide, judge, or commit through a simply conscious rational deliberation. It might appear that we are decisively and rationally deliberating as we consciously weigh different evidence for a view, as we entertain multiple possible commitments, or as we decide what judgment to render. But, the consequential judgments, decisions, and commitments take place behind the conscious scene, and we become aware of these decisions and judgements by observing our behavior generally and our subtle physical reactions to the unconsciously determined judgements and decisions, hence the "sensory" component of Carruthers' Interpretive Sensory Access (ISA).

Carrruthers is critical of some who use the system-one/system-two way of thinking to claim that the judgements and the decisions come from the rationally deliberative and conscious cognitive agency of system-two thinking of the individual. (Carruthers, 2007, section 3.2; 2011, 98). After reviewing the work of people who think the two-system approach works against his ISA approach, he rather argues it is more reasonable to think the two-system approach supports his ISA. He acknowledges the importance of system-two thinking. Conscious deliberation can contribute substantially to the final unconscious deliberation that decides the issue

as to what one believes. It is just that system-two conscious deliberation doesn't have the final say as to what one believes; that occurs in post conscious deliberation, as exemplified in the next paragraph.

Here is an example of what is going on from Carruthers' perspective. Suppose a mother has lost her child to a car accident instigated by a drunk driver (Carruthers, 2011, 152). Of course, she is angry because the drunk driver killed her son. The mother can know the "coarsely-individuated object" (152) of her anger. But exactly why is she angry? Is it the fact that her son is dead and not just injured that is the precise reason for her anger? Is it the fact that he was killed by a drunk driver that causes her dysphoria? Would she have been just as angry if the son had been killed rather by someone texting while driving? She can know reliably all the mental states exclusively generated from her sensory context, like the fact that she is mad at a particular person, the valence of the anger, and the object of her anger. But, the answer to these questions doesn't come through conscious rational deliberation. The judgment doesn't transparently show itself when settling the issue because these decisions are made at the unconscious level in post-conscious deliberation which decides finally if one is committed to a belief. It comes from unconscious motives, decisions, and judgments. The consequential judgments come ultimately from the unconscious level, not decisively from the agent's conscious rational deliberation, though the rational deliberation of the agent can play a causal role influencing how the decision is made in post-conscious deliberation. The specific reasons for, and response to, the anger or affect are determined "in mechanisms buried deep within the brain, utilizing inputs and decision criteria that are inaccessible to consciousness (Phan et al., 2002; Murphy et al., 2003; Schroeder, 2004; Barrett and Bar, 2009; Ochsner et al., 2009)" (Carruthers, 2011, 148). Substantial judgments and decisions are made in the deep inaccessible recesses of one's unconscious mind.

In summary, Carruthers asserts that individual agents simply don't have the kind of agency that Moran thinks they have, namely, an agency that consciously and finally deliberates about the reasons for believing p, an agency that knows its beliefs in the process of settling an issue, an agency that knows its mental states without observation. That agency, Carruthers asserts, given the evidence from self-observation and confabulation studies, simply doesn't exist. In no way is Carruthers saying that we can never know our substantial judgments,

decisions, and affect motives. Rather, people just can't know them by transparent rational deliberation (Carruthers, 2011, 152).

Chapter Two

Mindfulness Research Shows Limitations of Observational-Interpretive Models

Chapter Two argues that Carruthers and anyone who supports the never-any-privileged-access view have overgeneralized since mindfulness studies show that people often do have first person authority with adequate mindfulness, and this privileged self-knowledge is produced in a state observational-interpretive models say it can't be produced in, a state of direct and real-time consciousness of the mental state. Observational-interpretive models think self-knowledge is produced in a state without direct and real-time consciousness of the mental state mediated by interpretation, unconscious deliberation, or inference. From here on we will refer to "a state of direct and real-time consciousness of the mental state" as simply "conscious state."

"A state of direct and real-time consciousness of the mental state" = "conscious state"

The fact that privileged self-knowledge claims are initially generated in a conscious state fully aware of the mental states in real time as it is generated is one of the most important things that distinguish observational-interpretive and privileged access methods. Empirical research presents privileged self-knowledge that can't be produced by observational-interpretive access. There is no doubt that mindfulness is strongly correlated positively with privileged self-knowledge. The empirical studies we will point out bear this out. Thus, because observational-interpretive access can't produce privilege self-knowledge, models which are exclusively observational-interpretive oriented are limited in what they can account for within their model. Both the observational-interpretive and the privileged methods are needed for self-knowledge. In the process of establishing this conclusion, we see that each method is better differentiated and better appreciated for what it uniquely does with the help of mindfulness and empirical research on mindfulness. In reaching this conclusion, we have shown that each method is integrated with mindfulness. In the next chapter, we will use this finding of integration to talk about the limitations of privileged self-knowledge claims.

We want to know if the substantial characteristics of the privileged self-knowledge generated by mindfulness is produced by a conscious state or an unconscious state. This issue will be answered by comparing how likely it would be that an unconscious versus a conscious deliberative process could produce the

characteristics of the substantial self-knowledge proven to result from mindfulness. To get to the point where we can engage this crucial comparison, I first describe what it would take, according to Carruthers' ISA theory, for an unconscious deliberative process to produce substantive self-knowledge. I then describe substantial characteristics of the self-knowledge proven to be produced by mindfulness. Finally, I discuss what it would take for an unconscious deliberative process to produce the characteristics of the substantial self-knowledge resulting from mindfulness. At this point I will be able to compare how likely it would be that an unconscious versus a conscious deliberative process could produce the characteristic features of the self-knowledge produced by mindfulness. This will allow us to determine which process of deliberation is most likely to produce the unique characteristics of the self-knowledge inadvertently caused by mindfulness.

In coming to these conclusions we have made progress on our objective to derive the Assessment Framework. The Second Chapter develops three of the six key components of target disagreements that are used to derive the Assessment Framework, the observational-interpretive access method, the privileged access method, and mindfulness. Most importantly, the chapter demonstrates, with empirical research on mindfulness, that observational-interpretative access can't be the only way that we attain self-knowledge. Mindfulness studies show that there is privileged self-knowledge. And mindfulness shows how the two methods are interrelated.

What is mindfulness and how does it cause better self-knowledge?

What is mindfulness?

Let's start out with the best description and understanding of what mindfulness is. There is a consensus among scholars, psychologists, and practitioners that mindfulness has two crucial components: enhanced observation of one's surroundings and what is going on inside and outside of oneself, and a non-judgmental way of observing these things (Keng et alia., 2011). One of the most widely used definitions of mindfulness of scholars, psychologists, and practitioners focuses on these two things:

> "Paying attention in a particular way: on purpose, in the present moment, and nonjudgmentally" (Kabat-Zinn, 1994, p. 4).

That's it, those two things working together. In the process of looking at the studies below, we will get better, more complete understanding of what it is and how it works. In our daily lives we are mostly focused on goals, projects, jobs, relationships, childcare, etc., that take our attention, focus it, shape it, and guide it. Mindfulness facilitates a more complete and expansive awareness and observation. We notice our breath, heartbeat, micro-reactions to the things that are happening around us and inside of us. We notice how we interpret things around us, and the interpretations that we automatically put on our thoughts and everything we are experiencing. We notice the start and the ending of an emotion, an irritation, a desire, a trigger that influences us. We have a higher quantity of information about ourselves and things around us.

And the quality of the information we have about ourselves and our interactions with things and people is much increased by the second aspect of mindfulness, non-judgment. With the intense non-judgmental attitude by which we observe things, there is more balanced information about ourselves, our irritations, our triggers, how we interact with the environment and others. This is so because humans have a tendency to suppress negative information about themselves due to the unconscious self-enhancement motive, and we want to confirm the self-concept that we have about ourselves, and the non-judgmental attitude avoids much of the dysphoria and threat to self-concept that suppresses negative information. While this is the basic way that mindfulness works to produce indirectly so many health benefits and benefits for facilitating self-knowledge, we will see by looking at the following studies the fine-grained details of how mindfulness works.

Mindfulness isn't just a skill that Buddhists invented 2,500 years ago accessible only to those properly initiated. No, just the opposite; the research shows that all people to a higher or lower degree have the mindfulness trait; the results of the widespread use of a scale testing for mindfulness, the Five Facet Mindfulness Questionnaire, show that even people without any mindfulness training have this trait to some degree. Mindfulness as a disposition is believed to be normally distributed across individuals (Brown and Ryan, 2003).

Mindfulness inadvertently facilitates better self-knowledge

Before arguing that mindfulness fosters privileged access, we must first explain why any positive benefit mindfulness might have for self-knowledge can't come as a direct cause of mindfulness. Privileged self-knowledge can only be inadvertently fostered by mindfulness. The understanding of mindfulness that people

who promote it have, whether the person be the Buddha 2500 years ago or a practitioner or researcher today, is that mindfulness involves two, and only two, activities: enhanced awareness/observation with a non-judgmental attitude. That's it. Those two don't and can't produce privileged access to one's mental states by themselves, yet the research says they cause privileged access to occur. Their only remit is local, being aware of the here and now, of bodily functions, of micro things going on around one, and to do this without judgment about the things that show up. Because it doesn't judge ideas or beliefs that come to it, it can't possibly select for, or foster, the privileged access to mental states. Yet, and here is the most important point, in the presence of adequate mindfulness there is privileged access to one's mental states. How do we resolve this seeming contradiction? This situation can mean only one thing, that the causality of mindfulness is inadvertent, not directly expressed, not intended directly by it. Mindfulness doesn't tell the addict to stop eating sweets, stop using cocaine, stop hitting your spouse in a flurry of rage, stop ruminating, or love yourself. These judgments aren't part of mindfulness's remit. Any of the health and self-knowledge benefits of mindfulness have to be the inadvertent byproduct benefits of its passive remit.

Substantial self-knowledge, conscious or unconscious deliberation, mindfulness, privileged access

Unconsciously-produced, substantial self-knowledge

There is no doubt that mindfulness causes people to have self-knowledge. The empirical studies we will point out bear this out. The questions we ultimately want to answer: Does the substantial self-knowledge caused by mindfulness result initially from conscious processes aware in real time of the self-knowledge or unconscious deliberative processes?

ISA is committed to saying self-knowledge is formed in an unconscious context (Carruthers, 2011, 29). While Carruthers' ISA allows for some perceptually bound self-knowledge and judgements to be attained in a privileged, conscious, and transparent way, those are bound to a particular perceptual context, and so are not substantial (Carruthers, 2011, 159). Unlike perceptually bound self-knowledge, substantial judgments and self-knowledge are consequential for multiple contexts, occur separately downstream from perceptually bound judgments, and are generally the result of further reflection and inference (Carruthers, 2011, 75). Perceptually

bound judgments and self-knowledge have some of the same attributes that substantial judgments do; they can be decisions, or they can help one plan an event. For example, I see the person in the distance as my mother, or I hear the name called out as my own (Carruthers, 2011, 75). These judgments are bound to a perceptual context. They involve a minimal and quick understanding. They don't involve substantial deliberation. Notice the more substantial deliberation about the context in the following example from Carruthers of a substantial judgment: "I saw my mother come into the store, but it was so unexpected to see her there, in the capital city of a foreign country, that I did a double-take. But then I thought to myself, 'That really is Mother'" (Carruthers, 2011, 75). Here, the judgement confirms the initial sensory impression that it is mother. This involves intense reflection and inference. These judgments are perceptually grounded rather than perceptually embedded, says Carruthers (Carruthers, 2011, 75). They are formed downstream of the perceptual context. They are the final decision in a series of thoughts. They can be known only by interpretation, which is not transparent and not conscious (Carruthers, 2011, 75-6).

The first question will be answered by comparing how likely it would be that an unconscious versus a conscious deliberative process could produce the characteristics of the substantial self-knowledge proven to result from mindfulness. To get to the point where we can engage this crucial comparison, I first describe what it would take, according to Carruthers' ISA theory, for an unconscious deliberative process to produce substantive self-knowledge. I then describe substantial characteristics of the self-knowledge proven to be produced by mindfulness. Finally, I discuss what it would take for an unconscious deliberative process to produce the characteristics of the substantial self-knowledge resulting from mindfulness. At this point I will be able to compare how likely it would be that an unconscious versus a conscious deliberative process could produce the characteristic features of the self-knowledge produced by mindfulness. This will allow us to determine which process of deliberation is most likely to produce the unique characteristics of the self-knowledge inadvertently caused by mindfulness.

Before starting this comparative evaluation process, let's describe how, in the unconscious deliberation model of Carruthers' ISA, both unconscious and conscious systems can work together towards fostering accurate self-knowledge. There are many different unconscious processes with many different functions. There

is no one unified function of unconscious processes. They monitor systems, warn of danger, deliberate about what one believes, interrelate beliefs, and react to cues in the environment. The mindreading system can even monitor how well settled the beliefs are of the person just to make sure one hasn't made a mistake about what one believes. Internal speech is conscious, reflective, and, so, system-two reasoning; but it gets processed with unconscious system-one processes, which are quick, and don't use working memory (Carruthers, 2011, 58,101).

Fundamental to Carrruthers' ISA model is that an agent has real time conscious awareness of the sensory information which is broadcasted to the mindreading system and real time conscious awareness of any perceptually bound judgments; but one can't have real time conscious awareness of substantive decisions, judgments, and self-knowledge that finally settle an issue. Judgments and decisions are substantial when they can be used consciously in practical reasoning, when they are fully vetted in the unconscious context, when relevant considerations have been brought to bear, and when they have successfully satisfied all relevant unconscious demands and concerns. The content of the judgment or decision can be brought up consciously in inner speech. And inner speech can describe the aspects and consequences of a judgment. But, conscious deliberation can't consider the judgment as one's own final judgement until it has been endorsed by post conscious deliberation and subsequently presented to conscious awareness as a substantial judgement. Substantial decisions and judgments are the last and final say as to what one's decisions and judgments are; and they must be formed in an unconscious context, since only this context has all the information and processes need to determine what propositional attitude one has.

Conscious deliberation can inform the generation of substantial judgments and self-knowledge, but it can't decide on substantial judgements. The results of conscious thinking can be presented to the unconscious mindreading deliberation as proposed beliefs which may or may not be deemed substantial judgments of one's mental states. Conscious deliberation in system-two thinking takes place within a sea of system-one unconscious thinking culminating in a final and definitive unconscious deliberation determining what one's mental state is; and unconscious systems operate within a sea of conscious systems (Carruthers, 2011, 100). It might appear that one consciously and definitively deliberates, and through this process knows what one's mental state is; but that, for Carruthers, is an illusion, a habit of thinking that doesn't reflect the research on

confabulation and self-observation. While there is a great passage from a 2007 article that expresses how it is an illusion when people think they consciously form a substantial judgment as the result of their conscious deliberation, we will look at how Carruthers talks about this in a quote below from his 2011 book.

To attain the objectives mentioned at the beginning of this chapter, we must dig deeper into an understanding of the attributes and provenance of substantive beliefs, judgments, self-knowledge and decisions. We will see that the self-knowledge we will focus on soon, which empirical studies prove result inadvertently from mindfulness, have nearly all the substantive attributes the observational-interpretive model of Carruthers describes. The only question: Whether the provenance of those substantive beliefs caused by mindfulness is a conscious state with real time awareness of them or an unconscious state? Only after developing a careful understanding of the substantial attributes and their provenance according to his observational-interpretive and unconscious model will we be able to answer this question.

So, for Carruthers, there are many attributes that a judgment, decision, or belief about the self must have to be final and substantive, and they can only attain substantive status in the unconscious context. For example, one has to know if one really wants the judgment or the decision. For another, there has to be a decision about whether the proposed judgment or decision fits with all of one's other beliefs. As he says in the quote just below, for a decision to be substantial it has to interact "with an appropriate higher-order desire." Any conscious assertion that one will do something must be followed by an unconscious higher-order belief and goal (Carruthers, 2011, 97). As he also says below, they are formed downstream of the conscious activity; they are finally formed post-consciously. In other words, substantive judgements, decisions, and evaluation can't be formed in consciousness. Consider what Carruthers says in the following quote; which is rather long, but every part of it is used to gain understanding about the characteristics of unconscious deliberations for determining what mental state one has:

Decisions

Put differently, while a decision, if it is genuinely to count as such, can be followed by further deliberation, this should only be deliberation about the means to execute the action, not about the action itself. So if the act of buying a book is Q, the deliberation that follows a decision to do Q shouldn't be about whether or not to do Q (that should already have been settled), but merely about how to do Q in the circumstances.

In a case of System 2 decision-making, in contrast, the conscious event of saying to myself in inner speech, "I shall do Q," doesn't settle that I do Q, and the further (unconscious) practical reasoning that takes place prior to action is about whether or not to do Q. For on the account of System 2 practical reasoning sketched above, the sentence, "I shall do Q" (when heard

> as a decision to do Q, or as a commitment to do Q) only leads to the act of doing Q through its interaction with an appropriate higher-order desire (either to do what I have decided, or to execute my commitments). Thus the reasoning might proceed (unconsciously) like this: "I have decided to do Q. I want to be strong-willed. So I shall do Q." (Note that the final step, here, is itself a decision to do Q, albeit an unconscious one.) This should be sufficient to disqualify the conscious event in question from counting as a genuine decision, even though it does play a causal role in the production of the action. For the role in question isn't the right sort of role required of a decision. The real decision is undertaken unconsciously, downstream of the conscious event.
>
> **Judgements**
>
> Similar points hold with respect to judgments. A judgment that P should be apt to give rise to a stored belief that P immediately, without further judgment-related reasoning needing to occur. And a judgment that P should also be immediately and non-inferentially available to inform practical reasoning. Consider someone who wants Q, and who already believes that the truth of P would enable performance of an action that would bring about Q. Then, forming the judgment that P should be capable of interacting with the relevant belief and desire to issue in a decision to act. However, a System 2 "judgment" has none of these properties.
>
> Suppose that I say to myself, "P," and that (subsequent to the interpretive work of the mindreading faculty) this is heard as expressing a judgment that P, or as a commitment to the truth of P. This isn't by itself apt to give rise to a stored belief with the content P, but rather to the belief that I have judged that P, or to the belief that I have committed myself to the truth of P. And likewise, interactions with my other beliefs and goals will need to be mediated by a desire to behave consistently with what I believe myself to have judged, or by a desire to execute my commitments. These aren't the right kinds of causal roles required for an event to be a genuine judgment.
>
> In order to see the necessity of these constraints on what can count as a judgment, notice that without them judging would be in many ways no different from wondering. If a judgment could be the sort of thing that isn't apt to lead directly to a semantic or episodic memory with the same content, then in this respect it isn't distinguishable from wondering. Both would be attitudes that are directed towards truth, but neither would be apt to give rise to a belief in the content of the attitude. (Carruthers, 2011, 103-4)

This all means that there can't be a substantial decision or judgment formed in a conscious context, and this goes as well for a judgment, decision, or knowledge about oneself. To use Carruthers' analogy from the quote just above, there can't be a fully substantial judgment or decision formed upstream of the post-conscious deliberation. ISA rejects the idea that substantial judgments and decisions can be formed in a conscious context upstream of all the things needed for decisions and judgments to have substantial and consequential roles that are accepted as final. As he says just above, "The real decision is undertaken unconsciously, downstream of the conscious event."

One of the roles a judgment, say P, has when it is substantial is that it can "give rise to a stored belief that P immediately, without further judgment-related reasoning needing to occur" (from the quote just above). Later in the quote above he refers to this stored belief of P as "a semantic or episodic memory with the same content." This is an extremely important characteristic of a substantial judgement, since unconscious deliberation can't, by Carruthers' own admission, use working memory, and so can't hold many beliefs in short term memory. So, if unconscious beliefs are to be consulted during unconscious deliberation, they must be consulted one by one, not as a whole. Beliefs can't be interrelated in the unconscious state of mindreading (Carruthers, 2011). Conscious deliberation can hold many beliefs in working memory. Another role a substantial judgment has is

that it is "immediately and non-inferentially available to inform practical reasoning" (quoted above). Also, suppose a person wants something, Q, and she believes that the truth of a particular judgment, P, would enable the performance of an action that would bring about Q; if P is substantial and so the final word on the matter, then it is able to interact with the relevant belief and desire to issue in a decision to act. There is always a delay for being aware of substantive decisions and judgments because they have to be formed as the final word on the issue through an unconscious process before one can be aware of them. One can't be aware consciously of a judgement becoming substantial in real time—that is, one can't be conscious of the judgment at the same time that it takes on substantial attributes—because a judgment can only become substantial in an unconscious context. So, this means one can't have a real time awareness and differentiation of one's mental states.

Carruthers also gives an example (in the quote above) of the special role of a decision which is substantial. Let's call the action of buying a book Q. If a decision to do Q is substantial, the only deliberation that can follow it is how to do it, not whether to do Q. The characteristic of a substantial decision is that it decides the issue, that it commits to the action. There is no necessary commitment with inner speech. People wonder about doing many different things without committing to any one of the things we wonder about doing (Carruthers, 2011, 97). In inner speech we rehearse different possible actions without committing to any (Carruthers, 2011, 100). The imagistic representations of the rehearsals become inputs to the full suite of system one systems, like the mindreading system. These system one systems activate relevant memories and emotional reactions. And the affective consequences of the rehearsed actions are monitored by system one processes which feed back into the rehearsals in system two inner speech (Carruthers, 2011, 100); what we see from this is that the mental rehearsals of different possible actions which are globally broadcasted have unconscious cognitive activity immediately preceding and immediately following the broadcast images. Whatever is conscious is preceded and followed by a sea of unconscious processes. Mental conscious rehearsal is slower than system one because it co-opts the resources of the various system one processes (Carruthers, 2011, 100). Conscious activities are not operating alongside of system one activities; they are partly realized in cycles of operation of system one thinking (Carruthers, 2011, 100). And sometimes when one is consciously deliberating, the unconscious system one processes activate and give solutions to problems that are consciously rehearsed; there is, in this

way fluid and seamless interactions among the system one and system two processes (Carruthers, 2011, 100-101).

The fluidity and seamlessness of the interactions between system one and system two processes tell us why, according to Carruthers, scholars like Moran, and most people in the world, think they consciously come to know their own mental states when they are formed. He explicitly says he is "vindicating" some of the ideas of Moran (Carruthers, 2011, 101). People go about their lives consciously deliberating about issues around them. They are rehearsing consciously many different solutions. For example, I rehearse with images many solutions; and in one of those rehearsals, I say to myself what amounts to a conscious mental state, "I shall go to the bank." During the process of rehearsing one of the many different possibilities for resolving an issue, the unconscious system one process activates and endorses one of the rehearsed solutions with a higher-order motive (Carruthers, 2011, 116), like a standing normative belief "I should do what I have committed myself to doing" or "I should be a strong-willed person" (Carruthers, 2011, 101), or "I should financially be responsible for myself." The infusion of the system one unconscious endorsement is so swift that it appears falsely that one consciously came up with, and endorsed, the interpretation. In the example, I think falsely that I, by the mental state "I shall go to the bank," consciously decided the issue, when in reality the mental state made in the context of one imagistic rehearsal of a solution functioned as a cue activating an interpretation that was decided on in post-conscious deliberation. The final decision in unconscious deliberation downstream to endorse a particular conscious rehearsal's solution upstream is seamlessly transferred upstream so quickly that the person falsely thinks the conscious mental state expressing the rehearsed solution finally decided the issue when it really didn't. Instead, the individual consciously activates a particular interpretation made originally in post-conscious deliberation. In our example, the person falsely thinks the mental state expressed as "I shall go to the bank" decides the issue, when in reality the issue was decided in post-conscious deliberation. This is the basis of ISA's claim that nobody ever has direct access to their beliefs as attitudinal mental states in real time as they are formed. Following Carruthers' analogy**,** the conscious rehearsal's inner speech flows down the river seemingly deciding the issue in real time when in fact the issue was decided finally downstream and inserted upstream in the inner speech so quickly that it appears falsely to the person that they consciously came to the

conclusion. Some of Moran's views have been vindicated in the sense that Moran excellently describes well what merely seems to be happening, but really isn't.

The following passage best describes the seamless process described just above whereby one thinks falsely one is deciding an issue consciously yet in fact is adopting an already-available, sensorily-cued, interpretation already finally decided in unconscious deliberation:

> It is possible to claim, therefore, that transparent knowledge of our own attitudes exists at the System 2 level. And this would be consistent with the claim (p.102) that interpretation is ubiquitously involved in any episode of inner speech. For the interpretation of myself as deciding to go to the bank, or as committing myself to go, doesn't need to give me access to an independent event of the appropriate sort. Rather, the imagistic event comes to constitute an attitude of the kind in question. For it ensures that my subsequent thinking and acting will be just as if I had formed that attitude. Moreover, the interpretation occurs upstream of (and prior to) the globally broadcast imagistic event. As a result of interpretation, one hears oneself as making a commitment, or as expressing an intention or belief. The imagistic event thus embeds a higher-order judgment that one is making a commitment, or expressing an intention or belief. And it is because of this judgment, together with one's desire to execute one's commitments, or to act in ways consistent with one's attitudes, that the event in question comes to constitute the formation of a novel first-order attitude (Carruthers, 2011, 101-2).

The event of interpreting one's imagistic inner speech in which the results of an unconscious deliberation are embedded constitutes the attitudinal mental state of a belief. One doesn't have conscious access to the appropriate mental state directly.

Summarizing what we can say about a substantial judgment, decision, or instance of self-knowledge, they become substantial when they are:

Necessary attributes of substantiality of judgments, decisions, or self-knowledge

- **Final**: The decision or judgment is final, i.e., the issue is settled
- **Actionable**: They are "immediately and non-inferentially available to inform practical reasoning" (quoted above)
- **Committed**: There is a commitment to the judgement, decision, or belief
- **Higher-interacting**: The decision has to interact "with an appropriate higher-order desire" (quoted above)
- **Not wonderings/suppositions**: The judgments, decisions, and self-knowledge are not something that one merely wonders about.
- **Formed pre-consciously**: There can be no awareness of substantial judgements, decisions, and beliefs at the time they are initially formed. Awareness of them is always delayed because they have to be formed in an unconscious context and then one can become aware of them.

Without these essential attributes listed just above, any judgment, decision, or claim to self-knowledge is just what Carruthers calls a mere "wondering" or "supposition" (Carruthers, 2011, 104-6) and not a viable candidate for being a substantive belief. The job of the unconscious mindreading faculty is to form substantive beliefs and knowledge. I can wonder many things, whether I will have a heart attack soon, whether my son will be safe in his new school, whether Trump will be elected for a second term. But these wonderings aren't consequential, I don't commit to them, and they can't guide my actions. On the other hand, we live our lives

by substantial judgments and decisions and the beliefs and knowledge they yield. They are the beliefs that matter. And the attributes above give us the justification we need to know with confidence the contents of our attitudes, to know our mental states.

We can now summarize all the information the mindreading faculty has available for forming substantive beliefs about the self, judgments, and decisions:

The information the mindreading faculty has available for forming and initiate substantive beliefs:

Sensory observation inputs about oneself and one's environment

- Perceptions of one's own circumstances of environment and social context (Carruthers, 2011, 378),
- One's overt behavior (Carruthers, 2011, 378),
- One's affective feelings (Carruthers, 2011, 378),
- One's visual images (Carruthers, 2011, 369, 378),
- One's bodily experiences from interoception, proprioception, etc (52)
- Sensory events (Carruthers, 2011, 369)
- The contents of conscious working memory (Carruthers, 2011, 166)

Domain-specific memory based on sensory inputs

- Beliefs stored in memory formed originally in the unconscious deliberation
- Stored decisions about judgements of self and others 69 71

Imagistic information inputs from conscious deliberation and other mental activity

- Sentences in inner speech (Carruthers, 2011, 369, 378).
- Consciously-originated and non-substantial decisions, judgments, and their interactions only to the extent that they are rehearsed and engaged in inner speech and so "bound into" (58) perceptual representations, and only to the extent that they are presented in working memory and globally broadcasted. (Carruthers,2011, 53-56, 166).
- Working memory only takes in attitudes that are sensorily bound (Carruthers, 2011, 58).
- Queries of unconscious deliberation to find out more information from conscious awareness. 70

Notice that there are different categories of inputs to the mindreading system.

Carruthers gives four reasons why he thinks the mindreading system unconsciously makes decisions about what one believes based on sensory observational-interpretive inputs. Confabulation and self-perception research shows that decisions about what one believes are made by unconscious decisions to embody interpretations in sensory cues. Confabulation shows that we interpret mental states based on sensory things because we make the same mistakes as we make when attributing mental states to sensory things and others. When conscious decision goes out on its own and makes a decision because it doesn't have guidance from cues, or cues are ambiguous, it inevitably isn't reliable. Second, direct sensory input is more simple; it doesn't require a system for knowing one's own mental states other than the one developed for knowing the minds of others. Third, it makes sense from an evolutionary perspective. It was more important for reproductive success to develop the ability to know the mental states of others through observation and interpretation over one's own mental states. We therefore developed through adaptation the interpretive sensory access to others first, and

then we used the method of access to understand our own mental states. Also, considering conscious attitudes would be computationally intractable, says Carruthers (Carruthers, 2011, 55). Were the decisions made based on conscious attitudes the mindreading system would be overwhelmed relating, categorizing, indexing, and searching for the attitudes (Carruthers, 2011).

Before leaving this topic, we can list the motives we have discovered in Carruthers' work guiding the formation of substantive attitudes about the self in the unconscious mindreading process:

1) **Self-knowledge**. To help the self understand what its own mental states actually are such that one has beliefs and knowledge about oneself whose propositional content is true.
2) **Coherence**. To make sure any new beliefs are consistent with prior judgments (See the quote just above).
3) **Monitoring interpretations**. Monitor how well the beliefs are functioning in contexts in an ongoing fashion. (Carruthers, 2011, 101-2. 207-8)
4) **Dissonance reduction.** Produce false beliefs about the self if need be in order to reduce cognitive stress and dysphoria (Carruthers, 2011, 360-5)

Unconscious deliberation has all the information developed in conscious deliberation. Any deliberation in the conscious context is broadcasted to the unconscious context. Since any deliberation in inner sense is conscious, the unconscious system receives that information through the real time broadcast. The unconscious system gets all the sensory input coming from all the senses. When judgments or decisions are rehearsed or wondered in inner speech they too are received as input for the mindreading faculty. Unlike the conscious state, the unconscious deliberation can't hold recollection of mental events together in memory; the unconscious context has no working memory, as does the conscious context. The mindreading faculty can receive the information of conscious working memory so long as this information is presented in rehearsals, but can't receive it all at once (Carruthers, 2011, 327). Each memory can be stored individually in the mindreading system. To access the stored memories, it must access them one by one, not all at once.

To make sure any new beliefs are consistent with prior judgments, the mindreading faculty must check each new proposal of a belief against all the other beliefs it has. Because it lacks working memory, the mindreading faculty has to check the proposed new belief against all the other beliefs one by one.

Much of the activity of the mindreading faculty is engaged monitoring how the beliefs are functioning in their contexts of use. Here the role is to look for situations where the beliefs may be in new contexts where they no longer adequately point out to the person what their mental state truly is. For this goal the mindreading

faculty is monitoring all the inputs from the broadcast for any anomalies that indicate a belief isn't reliable for forming true beliefs about the self.

There is another goal of the unconscious processes, to reduce extreme stress and discomfort of the agent regardless of the reality of what one's mental state really is. Here is how Carruthers describes this motive on the unconscious level to reduces the debilitating effects of cognitive dissonance:

> It appears, then, that while mindreading is involved in the creation of dissonance effects, the latter can't simply be explained as resulting from confabulated attributions of attitudes to oneself. Rather, the core of the phenomenon is first-order. … Here is how the phenomenon works. One first performs an action that conflicts with some norm or value, say, or which one believes will have consequences that conflict with some norm or value. The mindreading system represents one as having freely chosen to perform this action for no sufficiently-justifying external reason. One's motivational systems then respond to the representation of oneself performing the action in this way by producing arousal and negative valence, sometimes issuing in more fine-grained emotions like guilt or disgust. Then "attitude changes" are behaviors that are undertaken in an attempt to manage one's own emotions. By expressing an attitude that is weaker than or contrary to the one giving rise to the negative affect, one attempts to make the latter go away. And indeed, the evidence suggests that such attempts are generally successful (Eagly and Chaiken, 1993; Elliot and Devine, 1994). (Carruthers, 2011, 360)

The mindreading process doesn't always have self-knowledge as its goal. When one performs an action that goes against one's values, the mindreading system may endorse a false belief in order to reduce the dysphoria resulting from the dissonance. This endorsement of a knowingly false belief doesn't have the truth of one's mental states as a goal, but rather the survival of the individual as the goal.

Characteristics of substantial self-knowledge produced by mindfulness

So, let's proceed with the strategy we talked about earlier. We have seen what it takes, according to Carruthers' ISA theory, for an unconscious deliberative process to produce substantive self-knowledge. We now describe the unique characteristics of the mindfulness-produced self-knowledge with an eye ultimately to determining which of the two models, the conscious or the unconscious one, is more likely to produce the unique characteristics of the self-knowledge caused inadvertently by mindfulness. After this I will discuss what it would take for an unconscious deliberative process to produce the characteristic self-knowledge resulting from mindfulness. Just below is a list of the types of self-knowledge mindfulness has been proven to cause according to the empirical research. The results of these studies will be described in more detail farther down below. The following is a list of particular types of self-knowledge mindfulness causes, and we will pull examples from this list.

Particular types of self-knowledge empirical studies show mindfulness causes:

Mindfulness cause people to:

- Know better their self-concept, (Vago, 2014)
- Have more self-concept clarity, (Vago, 2014)
- Know better what in the future would reflect their self-concept. (Vago, 2014)
- Know what will make them happy in the future, (Vago, 2014)
- Know better their self-concept, (Hanley, 2017)
- Have greater clarity about one's beliefs, (Hanley, 2017)
- Make better decisions that reflect one's self-concept, (Hanley, 2017)
- Have less ambiguity about their propositional attitudes, (Dummel, 2018)
- Know better how they would respond to emotional events, (Emanuel et al., 2010)
- Know better their feelings of self-worth, Koole et al., 2009)
- Know better their cravings, (Papies, Barsalou, an Custers, 2012)
- Know the cues and interpretations of their addiction, (Papies, Barsalou, and Custers, 2012)
- Know one's motives, (Kernis and Goldman 2006)
- Know different ways of relating to the dysphoric thoughts, (Ma and Teasdale 2004)
- Differentiate their discrete emotional experiences with less emotional difficulties and more effective emotional regulation, Hill and Updegraff, 2012)
- Remember with enhanced detail dysphoric memories, about their previous episodes of depression, (Hargus et al., 2010)
- Recognize current mental states, (Modinos, Ormel, and Aleman, 2010, 369)
- More aware of repetitive thoughts, (Feldman, Greeson, and Senville 2010, 1008)
- Understand negative dysphoric thoughts aren't necessarily representative of reality, (Perona-Garcelán et al. 2014)
- One has real time awareness and self-knowledge of one's mental state, (Tapper, 2018).

The type of self-knowledge that stands out as most appropriate to consider for our purposes follows:

One has real time awareness and self-knowledge of one's mental state.

Before arguing that this type of self-knowledge facilitated by mindfulness can't be explained through ISA, I must first explain how the research suggests the health benefits of mindfulness come about largely through the privileged self-knowledge that mindfulness fosters. To do this, I show how empirical studies indicate that mindfulness improves eating habits through self-knowledge. In the last section of this chapter, I will describe studies confirming that such health benefits largely come through the privileged self-knowledge mindfulness fosters.

In a close reading of empirical studies on the health benefits of mindfulness for reducing emotional eating, we can see how privileged self-knowledge plays a crucial role facilitating the health benefits. In a meta-analysis of 74 independent empirical studies on the health benefits of mindfulness for reducing emotional eating, Margarita Sala and others find that the studies point to the self-knowledge of mental states facilitated by mindfulness as a driving force for improving eating habits. The more mindfulness the more subjects can know and distance themselves from the eating-disorder (ED) thoughts triggered automatically by the food cues in

their environment (Sala et al., 2020, 848), e.g., ED thoughts like "I must eat in order to avoid emotional problems." And the more mindfulness the more subjects can distance themselves from the behaviors that the thoughts encourage, thus "decoupling the link between ED thoughts and behaviors" (Sala et al., 2020, 848). The enhanced awareness of mindfulness helps one see cues and triggers for what they are, and this reflection produces the distance that leads to "the ability to change one's perception and act according to personal values, even in challenging situations" (Sala, 2020, 847); in this quote Sala is paraphrasing the conclusions of another study (Hayes, Luoma, Bond, Masuda, & Lillis, 2006). This reflective distance leads to a change of perception whereby one can act on personal values, for example a value expressed like "I don't need to eat in order to avoid emotional problems." The interruption gives the subjects the reflective distance needed from the debilitating and automatic interpretation of the cue, freeing the subjects up to choose different beliefs and behaviors more conducive for healthy eating. Researchers agree that mindfulness reduces habitual emotional eating by interrupting automated interpretations of cues in the subjects' environments (Sala and all, 2019, 846). Other studies corroborate Sala's findings about reflective distance and perspective change (Lattimore, 2020, 650-654), (Levoy, 2017, 124), (Hsu, 2021, 2), (Verrier and Day, 2021, 104), and (Morillo-Sarto, 2023, 315).

Cues can be observed in internal events such as thoughts or feelings; they can be behaviors, particular contexts, inner speech, or objects. When triggered, the cues initiate an interpretation of one's beliefs, one's propositional attitudes. Those automated interpretations lead one to think one has mental states they may not have. By seeing the automated interpretations of cues <u>as</u> interpretations, mindfulness interrupts the automation. The unconscious automation is only interrupted by the conscious state of mindfulness; of this, the research is clear. And that interruption is done in real time, at the time of the conscious perception of the cue in the perceptual context. What is common among all of these instances of self-knowledge in the studies is that there is known to be an interruption of automated interpretations.

Unconscious provenance of mindfulness-produced self-knowledge?

The following is what ISA, or any other observational-interpretive/unconscious model for self-knowledge, has to show: How the self-knowledge resulting from mindfulness attained its substantiality in an unconscious state as the result of interpretation and sensory observation. As we have seen, the reflective distance provides

a context in which the subject can perceive the situation in a different way. Subjects come to see that the belief "I must eat in order to avoid emotional problems" is false. This judgment that mindfulness facilitates decides the issue in a final way. It isn't just a perceptually bound judgment, since it is making a judgment about the falsity of a longstanding belief that the person has to eat in order to avoid emotional problems. ISA has to be able to explain how the substantial judgment inadvertently caused by mindfulness originated from an unconscious context through interpretation and sensory observation.

On the face of it, it seems that the interruption described above and the self-knowledge following in its wake aren't the result of interpretation and sensory observation. The researchers of mindfulness listed above assume the agent consciously deliberated about the interruption and came to decide that she doesn't necessarily have to eat in order to avoid emotional problems. Agents don't yield to unconscious-deliberated judgments; they deliberate consciously themselves and finally decide the issue about what to make of the interruption experienced through mindfulness. The agent of her own accord consciously rejects the interpretation cued automatically to be activated—in the case of Sala et al. study, the interpretation that one must eat in order to avoid emotional problems. The agent deliberates consciously about the interpretation and decides that the interpretation of her mental state doesn't necessarily actually reflect her mental state. Mindfulness itself can't produce directly the belief that one is not necessarily needing to eat to avoid emotional problems because its only remit is to observe intensely without judgment, and such a self-belief is a judgment. Mindfulness, just by intensely pointing out all the details of one's perceptions, interoceptions, and environment, makes one able to see any interpretation <u>as</u> an interpretation. And this "seeing <u>as</u>" gives conscious deliberation just enough reflective distance from the interpretation to consciously assert its own belief about the self; such is the general view of the researchers of mindfulness.

As discussed, Carruthers would agree that the self-knowledge resulting from mindfulness appears to be uninterpreted and non-observational. He just thinks that the unconscious deliberation is so quickly and seamlessly integrated into conscious life that people falsely think it originated in conscious deliberation. This means the rejection of the automated interpretation described in the Sala et al. study and the other studies described above has to be initiated in post-conscious deliberation even though it really does appear to be

consciously generated by the agent. While the rejection may have taken place first in conscious rehearsal, a belief is final and takes on substantive status only in the unconscious deliberation context. When the post-conscious deliberation endorses the proposal, it becomes a substantial belief and has the power to inspire practical action and inferences based on it.

The information problem and skill problem

There are three problems with the unconscious models' explanation above of how the substantive self-knowledge is caused by mindfulness. First, there is an information problem. Unconscious models, at least Carruthers', can't adequately show, on their own terms, how there is access to the information that is needed to be able to form the distinctive self-knowledge known to occur as a result of mindfulness.

To be sure, the mindreading system has much information. Just look at the illustration above entitled "The information the mindreading faculty has available." The mindreading system has diverse sensory observational-interpretive information of oneself and one's environment; for example, we have information from proprioception, interoception, one's own overt behavior, one's feelings. It has different types of domain-specific memory, like beliefs originally from the unconscious process, and judgments of self and others. And we have many varieties of imagistic information inputs, the contents of inner speech, non-substantial judgments to the extent that they are rehearsed in inner speech, etc. It can have the results of its own urgent requests for conscious investigation of mindfulness, results which it would receive when there are conscious rehearsals and inner speech related to the request. It can request conscious deliberation to run rehearsals of particular issues it needs to understand better. Also, from conscious rehearsals and inner speech, unconscious deliberation can get positive impressions from empirical research of the statistical and empirical efficacy of mindfulness for facilitating reliable beliefs about the self. From those two conscious sources it can see that the statistical studies are based on actually observed sensory data coming from a huge variety of people who are participants of the research. It can even have some real life, and not just rehearsals, sensory information about how the individual has improved self-knowledge about minor things bound to a particular context. Conscious deliberation can yield privileged self-knowledge about the effects of mindfulness on the self so long as that self-knowledge is non-substantial and perceptually bound; for example, one is better at judging the person from afar as one's mother,

the individual is able to pay attention better at work, or the person can tolerate a little more negative information that she receives from others.

But the mindreading process doesn't have the most crucial information needed, information about how its own person tolerates the replacement of the substantial belief that unconscious deliberation put in place for practical reasons and that has for a long time been continuously re-endorsed. In the case we are considering, unconscious deliberation is being asked through conscious rehearsal to decide to endorse a new substantial belief, "I don't need to eat to avoid emotion problems," that replaces the long standing and equally long endorsed substantial belief, "I do need to eat to avoid emotional problems." And it can't have information about how mindfulness, with its known ability to help people tolerate dysphoria, helps the mindreading system's own person handle the loss of a problematic, yet substantial and practical, belief. To be sure, one can have information about how the individual with mindfulness tolerates better a little negative information related to perceptually bound beliefs about the self, and it can know the individual pays better attention at work. Those things can be managed through perceptual bounded judgments, decisions, and self-knowledge without having to be endorsed by unconscious deliberation. However, a very important and consequential belief could only be replaced by unconscious deliberation. For the agents themselves of the Sala study and others the unconscious deliberation doesn't have any information about how the individual would tolerate the replacement of a substantial and practical belief that is longstanding and continuously endorsed by unconscious deliberation. It doesn't have observational-interpretive information from past similar situations, because there aren't any, at least initially, examples of how the substantial belief change would be tolerated. The new replacement belief pointed to in the Sala study and others would be the first significant new belief replacement resulting from mindfulness. The interpretation triggered to occur with the cue and now interrupted by mindfulness observation is long standing. It has worked for the individual to get along in the world. And there is no information that indicates the new substantial belief would work better. Unconscious deliberation is largely set up to make decisions based on the inputs from sensory observation and from observation of events, as can be seen by the list above of things that it has access to.

Not having concrete and specific real life sensory information of how well the mindreading system's individual tolerates such substantial and extremely consequential belief replacements is crippling for unconscious deliberation. It doesn't have the ability and skill to weigh all together the merits of different possible solutions to the problem. It just isn't set up to do this, since it doesn't have a working memory of its own. It also isn't set up to consider all together hypothetical scenarios like the way conscious deliberation can. It does have access to the working memory of the conscious context, but only to the extent that this working memory has sensory information in it, and only to the extent that the conscious context rehearses what it has in its working memory. And even with access to the working memory of conscious deliberation, it has to store information in its domain specific storage, and it has to consider the stored information and beliefs one by one without holding all the information in memory all at once for evaluation. Also, it only gets information about the effectiveness of mindfulness for tolerating new beliefs after replacing longstanding substantial beliefs; at least initially, it doesn't have any of this information. Unconscious deliberation just doesn't have the ability to evaluate a belief about how well its individual would fair with mindfulness to the replacement of a longstanding substantial belief since the evaluation would have to be made without concrete relevant sensory information. That kind of evaluation just isn't part of its remit. One of the two things that ISA needs for unconscious deliberation is sensory information for how substantial beliefs would work, and it doesn't have it.

So maybe the unconscious deliberation about replacing a substantial belief can just rely on the conscious deliberation with its ability to hold in working memory all the different outcomes of how mindfulness could help one tolerate substantial belief replacement, and with its ability to take into account what all the statistical empirical mindfulness studies say about how it has helped others tolerate dysphoria. After all, it can do this successfully for perceptually bound beliefs.

Yet unconscious deliberation can't turn to conscious deliberation for help because conscious deliberation isn't reliable for reflecting what one's mental states actually are, and conscious deliberation doesn't know all the reasons why the false belief proposed to be replaced was put in place and consciously endorsed by unconscious deliberation. It is exactly because of the unreliability of conscious deliberation that Carruthers came to find that we don't have privileged access to our mental states about ourselves. Confabulation studies

show that when the determination of what belief the individual has is solely dependent on conscious deliberation, the individual is frequently wrong. One of the motives of unconscious deliberation is the truth of one's mental states, as we have seen. Unconscious deliberation generally doesn't trust conscious proposals in rehearsals for substantial beliefs. Another one of the motives of unconscious deliberation, as we have seen, is survival, as we have seen. While unconscious deliberation prefers its individual to have correct beliefs about the self, survival trumps the truth about the self when a false belief is put in place by unconscious deliberation to be triggered by the individual. Unconscious deliberation has the big picture, a desire for the truth of the mental states one has, but also a desire for the individual to survive. Conscious deliberation doesn't have this big picture. All in all, unconscious deliberation, as limited as it is, is more reliable for getting at the truth of one's mental states; and only when the truth about one's mental states can't be tolerated, as a last resort, it puts in place false beliefs that may produce dysphoria, but not the kind of dysphoria that makes the person depressed or unable to function. For all these reasons, the information problem definitively indicates that the particular self-knowledge produced through mindfulness, even on the strict terms of ISA, just can't be produced in unconscious deliberation.

The dysphoria problem

The substantial self-knowledge produced by mindfulness also can't originate in unconscious deliberation because there is a significant chance that the replacement of the longstanding substantial belief will precipitate fatal dysphoria. As we have seen, Carruthers thinks the mindreading system sometimes has the survival and mental health of its individual as a goal rather than the truth of its mental states. In the Sala study the subjects reject a belief about the self that is longstanding, has caused dysphoria for them, and has been determined to be false, namely, the belief that they necessarily need to eat in order to avoid emotional problems. From Carruthers' ISA perspective, the fact that the belief has been longstanding indicates that it has been continuously endorsed by the mindreading system and that it has survived many cycles of the monitoring that the mindreading faculty engages in order to make sure that beliefs are functioning. According to the passage on cognitive dissonance quoted above, (Carruthers, 2011, 360), all these things can mean only one thing: With the lesser

dysphoria of the longstanding false belief, the mindreading system is protecting its individual from a much more severe, and potentially harmful, dysphoria-producing true belief about the self.

Now we can appreciate the dire situation the mindreading faculty is in when there is an urgent request through conscious rehearsal to replace the longstanding belief with the new belief now able to be considered with the interruption due to mindfulness. To get rid of this longstanding belief about the self could bring back the severe, and potentially life threatening, dysphoria the false belief was put in place to reduce. The stakes are very high, and, as we have just proven above, the mindreading system doesn't have the type of information it needs in order to endorse the replacement belief.

Given all that has been said about this dysphoria situation, it appears that non-endorsement of the urgently proposed belief replacement is the only way to go for unconscious deliberation. It doesn't have any sensory information that supports the idea that the individual will be okay after belief replacement. In fact, the only sensory information that it has about this matter goes against the proposed replacement belief. The only sensory information it has is the stored memories of the horrible and life-threatening dysphoria that results when the individual doesn't have the belief. To be sure, the unconscious deliberation faculty has, through conscious rehearsals, statistical data from empirical research on mindfulness that it reduces dysphoria. While it has statistical and hypothetical evidence that the individual would be okay after belief replacement, most of the information the unconscious mindreading faculty relies on comes from sensory information about the individual's body, thoughts, and feelings. It deliberates over sensory information about the individual. The longstanding false belief creates some dysphoria, but the dysphoria produced when it is absent is orders of magnitude worse. The only safe thing to do for the unconscious deliberation faculty is to deny the request for belief replacement. The dysphoria problem also indicates the belief replacement isn't made in an unconscious state.

The assurance problem

A third problem the unconscious deliberation has with endorsing the belief replacement is that it can't take comfort in the assurance mindfulness studies provide that mindfulness helps people tolerate the dysphoria that results from accepting the truth about oneself. The assurance we are talking about here comes from the

understanding that mindfulness has been shown to help people tolerate dysphoria. Were the unconscious deliberation able to take confidence in the fact that mindfulness is believed to help people tolerate dysphoria, this would assure the unconscious deliberation that the individual will be okay if the longstanding belief is replaced.

Mindfulness has been shown by several statistical studies to be successful for lowering the dysphoria of others due to its nonjudgmental stance. Yet there is no adequate assurance that mindfulness will ameliorate the dysphoria with its non-judgmental stance in this particular situation with this particular individual. After all, the individual could be an outlier with regard to the statistical studies. There could be something about the individual that is idiosyncratic and not controlled for in the statistical studies. There could be something in the individual's context that isn't accounted for in the studies. And again, the unconscious deliberation is largely oriented around making decisions based on the observable sensory information about what is going on with the individual, not around judging the merits and relevance of statistical studies and mindfulness' way of relieving dysphoria.

Unconscious deliberation can't get assurance from perceptually bound mindfulness because this isn't dealing with the very consequential beliefs that are substantial. Even if mindfulness has been successfully tested through perceptually bound conscious deliberation, that isn't evidence for what would happen in the replacement of a substantive belief. In the rehearsals generated by conscious deliberation, the person is fine after belief replacement. But the unconscious deliberation can't trust hypothetical information and mere simulations.

All three of these problems individually, but especially all together, prove deadly for any chance that unconscious deliberation could have produced the substantial self-knowledge proven to be caused by mindfulness. The unconscious deliberation of the mindreading faculty just isn't in the business of making decisions about substantial beliefs based on rehearsals of hypothetical and statistical information. It doesn't have the working memory needed to compare the merits of solutions side by side. It is built to determine what substantial beliefs to put in place based primarily on the information presented to it about its individual's proprioception, interoception, context of the event, coherence with other substantial beliefs.

While the major goal of unconscious deliberation is to get at the truth of the matter as to what one's belief is and what one's mental state is, a second goal is to keep its individual alive and healthy, physically and mentally. And when those two goals are at odds, when severe dysphoria brings into question its individual's health, the unconscious deliberation faculty will choose the second goal over the first; it will choose the survival of the individual over the truth of the person's mental state and belief. Also, with such high stakes, it can't trust the research on mindfulness that says there is a good chance that mindfulness reduces the dysphoria of troubling beliefs such that it doesn't have to install false and substantive beliefs, and such that it can replace a false belief with one that reflects more the person's belief and mental state.

All this means that the substantial self-knowledge proven to be produced by mindfulness can't be explained on the ISA model. While unconscious deliberation certainly determines much of our substantive self-knowledge, it just as certainly doesn't determine some substantive self-knowledge. Carruthers is wrong to think that all substantive self-knowledge is produced through unconscious deliberation.

Answering the first question: Produced in a conscious state or unconscious state?

Now we can answer the first question posed in this chapter: Does the substantial self-knowledge caused by mindfulness result initially from conscious or unconscious deliberative processes? The answer is that the self-beliefs resulting from mindfulness don't get their substantiality in an unconscious state; rather they get it from a conscious state aware of the substantial self-beliefs in real time.

While we have seen unconscious deliberation doesn't have the cognitive skills and remit needed to successfully deliberate and endorse a mindfulness-produced belief about the self replacing a false but extremely effective belief, conscious deliberation does. We have seen that unconscious deliberation isn't able to hold many possible solutions to a problem in working memory; it can't even rehearse possible solutions or ponder hypothetical situations one can see oneself in. It can't seriously consider the results of empirical and statistical research on mindfulness, since the sensory information the research is based on came from other people, and unconscious deliberation mostly relies on sensory information from the individual's behavior, speech, surrounding environment, and context one is in. It doesn't trust empirical research as much as it trusts sensory information from one's self and from one's environment. It doesn't have the information it needs. It has an

information problem. After all, unconscious deliberation can't know that its individual isn't just an outlier, just one of the few individuals who doesn't experience the relief from dysphoria that others do in the mindfulness condition. Or perhaps one's individual is significantly unlike all the other participants in the research on mindfulness such that the results of mindfulness studies can't say what is likely to occur for one's individual in the mindfulness condition. So, it can't get any assurance from the empirical studies that are optimistic about subjects being able to deal with the dysphoria that results from such belief replacement. It has an assurance problem. It also has a dysphoria problem. The stakes are high because the dysphoria is severe. With such high dysphoria stakes, without adequate information, and without the assurance of the mindfulness research, the unconscious deliberation of mindreading is crippled.

Conscious deliberation doesn't have these problems. It has the information it needs to make the kind of decision needed for the substantial self-knowledge inadvertently produced by mindfulness. It has information from holding different solutions in working memory in order to judge their merits. It has information from its detailed rehearsals of how one would deal with dysphoria in different situations and with different solutions to the problem.

Conscious deliberation doesn't have the assurance problem. It can take solace in the assurance given by mindfulness studies which say there is a very good chance that the dysphoria will be manageable because of the nonjudgmental awareness that mindfulness encourages. It can compare in working memory the detailed attributes of the subjects taking part in the empirical mindfulness studies to the detailed attributes of one's self in order to decide how likely one would have the same results as the majority of people do in the studies. It can in this way gauge how likely it is that one would be an outlier in the studies. Because it has these skills, it isn't crippled by the fact that there is no concrete sensory information as to how one would fair when replacing the false belief. It can make a final decision on false belief replacement without actual sensory information about how one would fair without the false belief.

Conscious deliberation isn't crippled by the high-stakes dysphoria problem. Conscious deliberation can weigh the risks of belief replacement of a false belief against the risks of keeping the status quo. It has the information, working memory, and skills needed to gauge how well one's own situation is enough like the

situations of the subjects in the empirical studies to take assurance that the research yields that the dysphoria is manageable in the mindfulness condition.

For all these reasons, the answer to the question of this chapter is clear. The substantial self-knowledge caused by mindfulness results initially from conscious deliberation with real time awareness of the self-knowledge. Only conscious deliberation has the ability to endorse the false belief replacement. The final decision made to embrace the substantive self-knowledge has to have been made in conscious deliberation without the sensory information that unconscious deliberation needs. The final decision isn't interpreted in the sense that it isn't being told what belief one has by unconscious deliberation. Conscious deliberation can make rehearsals that include the new replacement belief. It can work with statistical principles. It can gauge the relationship between oneself and the studies done with the sensory information of other participants in the study. The fact that privileged self-knowledge claims are initially generated from a conscious state and are aware of mental states at the time of their generation is one of the most important things that differentiates observational-interpretive and privileged access methods.

In the process of answering with the help of mindfulness this question about the conscious state origin of the privileged self-knowledge produced by mindfulness, we find that each method can be appreciated for what it uniquely does. In Part One and in this chapter, we have pointed out the need for observational-interpretive access in our daily lives. Mindfulness, as we have said, just shows interpretations as interpretations. While mindfulness doesn't judge interpretations, it does give one reflective distance to be able to judge them. As we have seen, some interpretations, like "I need to eat in order to avoid emotional," are sometimes consciously replaced. But we also have found that the research shows we need some automated beliefs from the observational-interpretive method to live our everyday lives effectively. Both methods can be appreciated for what they uniquely do.

Health benefits through fostering privileged self-knowledge

I present here a detailed overview of two independent studies that corroborate my earlier conclusion that the self-knowledge facilitated by mindfulness is produced consciously in a privileged way and leads to health benefits. This earlier conclusion was attained by closely reading the empirical studies that positively correlate

mindfulness directly with health benefits, and by finding in these studies the feature of mindfulness that most facilitates the health benefits. That key feature is the reflective distance largely facilitated by the non-judgmental facet of mindfulness. That feature, as we have seen, inadvertently creates a context in which the self can come to know something about itself free of the automated responses to cues that beholden it to the self-beliefs leading to unhealthy habits. Thus freed up, the person can consciously deliberate forming privileged self-knowledge that promotes health benefits. In effect, I found evidence in these earlier studies that privileged self-knowledge mediates indirectly the relationship between mindfulness and health benefits.

Corroborating my earlier finding, I now present the empirical studies of Maryam Abbasi and Adam Hanley which both conclude that there are two pathways from the effects of mindfulness to the health benefits, one directly from mindfulness to the health benefits and one indirectly from the privileged self-knowledge facilitated by mindfulness to the health benefits (Abbasi et al., 2020) (Hanley et al., 2017). These two mediation studies use different methods, different approaches to mindfulness (state versus disposition), and take place in very different cultural contexts with different pools of participants. Yet they both come to the same conclusion mentioned just above. Erika Carlson does a good job of summarizing the empirical evidence for thinking that self-knowledge mediates the relationship between mindfulness and health benefits (Carlson, 2013).

Maryam Abbasi's mediation study approaches mindfulness as a state. In a state study, one looks at how participants feel or act regarding mindfulness in a specific moment based on their current situation. In a disposition study, one looks at how participants generally feel or act regarding mindfulness based on their consistent and overall character. Consequently, Abbasi's study is designed first to measure before an intervention the initial state of participants' mindfulness and the other two variables—i.e., self-knowledge and health benefits. After the intervention, the same three variables are measured to see how the intervention influenced the subjects' state.

I give the details here of Abbasi's study so as to compare them later to those of Hanley's. 118 Iranians were recruited who have stress-related symptoms. The participants went through an 8-week mindfulness program called Mindfulness-Based Stress Reduction (MBSR), a program developed by Kabat-Zinn (Kabat-Zinn 1982). The instructor was certified to teach MBSR. MBSR was administered for 3 hours weekly in a

group format for 8 weeks. In each section participants learned about psychological and physical effects of stress, and were taught mindfulness practices, like the awareness of body, breath, sounds, thoughts, and emotions; attention to the body sensations from toes to head; yoga body postures and practices designed to improve awareness of the musculoskeletal system and to strengthen its balance; mindful eating, walking, and talking. Participants were instructed to practice the techniques at home for 30 minutes. Three surveys are used before and after in order to determine changes of the three variables due to the intervention: Stress is measured by the Depression, Anxiety, Stress (DASS); mindfulness by the Five-Faceted Mindfulness Questionnaire (FFMQ, and self-knowledge by the Integrative Self-Knowledge Scale (ISK). The ISK scale includes 12 items that record the efforts of individuals to synthesize past, present, and desired future self-experiences into a meaningful whole (Ghorbani et al., 2008). For example, "If I need to, I can reflect on myself and clearly understand the feelings and attitudes behind my past behaviors." It was hypothesized that improvements in the facet of mindfulness called non-judging awareness would positively affect changes in psychological symptoms indirectly through positive changes in self-knowledge at post-treatment. Abbasi's study engages a bootstrapping method.

The results of Abbasi's state study clearly support the understanding that privileged self-knowledge mediates the effects of mindfulness facilitating psychological well-being, and we will soon see how the results of Hanley's disposition study are remarkably similar. Mediation analyses revealed that changes in self-knowledge significantly mediated the relationship between changes in non-judgmental awareness and well-being. The following is the indirect effect due to self-knowledge: $\beta = 0.12$, 95% CI [0.03, 0.27]. The indirect effect of mindfulness on well-being through self-knowledge is .12, and there is 95% confident that the true indirect effect in the population lies somewhere between .03 and .27. β represents the change in the dependent variable for a one standard deviation change in the independent variable. Changes in self-knowledge significantly mediate the relationship between mindfulness and psychological symptoms. Since non-judging is the most important facet for promoting self-knowledge, and since reflective distance is highly correlated with non-judging, this makes reflective distance a key factor in the facilitation of self-knowledge.

Hanley's very different approach to mindfulness as a disposition influences the design of his mediation study. Because he is focused on a relatively stable dispositional trait rather than a state that can be induced or cultivated by, for example, a program, there is no need for an intervention. Consequently, the measurement of the three variables need only be taken once. The goal of the study is to understand the inherent levels of dispositional mindfulness present in participants, and to assess how the different levels contribute to self-knowledge and psychological well-being.

For the background of the study, there were 1089 university students participating from a large university in the Southeast of the United States. Dispositional mindfulness (DM) was measured with the Five Facet Mindfulness Questionnaire (Baer et al., 2006). Non-judging awareness is one of the five facets in this questionnaire (e.g., "I make judgments about whether my thoughts are good or bad"). Self-concept clarity was measured with the Self-Concept Clarity Scale (SCC) (Campbell et al., 1996), (e.g., "In general, I have a clear sense of who I am and what I am"). Psychological well-being was measured with the Scales of Psychological Well-Being short form (PWB) (Ryff & Keyes, 1995). This study employed structural equation modeling (SEM) to explore the extent to which SCC mediates the relationship between DM and PWB.

Like Abbasi's study, Hanley's concludes that self-knowledge mediates significantly the relationship between mindfulness and psychological well-being. The non-judgmental facet of mindfulness was determined to have the strongest influence on self-knowledge ($B = 0.36$, $p < 0.001$). "$B = 0.36$" represents the beta coefficient for the non-judging facet in the regression structural equation model. And the small p-value indicates the relationship is statistically reliable. Together they indicate a positive correlation between the non-judging facet of mindfulness and self-knowledge. And the self-knowledge facilitated by the non-judging facet significantly contributes to psychological well-being through its indirect influence. Since the study concludes that self-knowledge mediates the relationship between the non-judging facet and psychological well-being, it is reasonable to conclude that the non-judging facet significantly contributes to psychological well-being through its influence on self-knowledge. And given that the non-judging facet has the strongest association with self-knowledge, it is also reasonable to suspect that the non-judging facet has the strongest association, mediated through self-knowledge, with psychological well-being.

The significance of these studies and the ones like it is huge for what was said earlier in this chapter and for the Assessment Framework developed later on. The results confirm what we inferred from a close reading of the Sala et al. study. In the studies focused on earlier, the feature of mindfulness that most promotes the health benefits is the reflective distance which itself is facilitated by the non-judging facet. The studies of Abbasi and Hanley both confirm that the non-judging awareness of mindfulness is strongly correlated with the self-knowledge that mediates the health benefits of mindfulness. Non-judging awareness works to facilitate the reflective distance that provides a bias-free context for a free understanding of the self that facilitates the health benefits.

These two studies support the earlier asserted claim that the self-knowledge facilitated by mindfulness is made in a conscious context. Both studies assert that reflective distance fostered by non-judging is a key feature. Reflective distance requires consciousness. Mindfulness, as discussed earlier, can't produce the self-knowledge on its own. The interruption gives the subjects the reflective distance from the automated interpretation of the cue needed in order to choose different beliefs or behaviors. The two studies prove that the reflective distance of non-judging leads to self-knowledge in a conscious context. Only in a conscious context can the non-biased and free deliberation take place.

The fact that these similar results are made from very different contexts lends support for my view. The Abbasi student had students from Iran, while Hanley's study is composed of students from the South-East United States. The Abassi study is a state study, while the Hanley's is a disposition study. Abassi has an intervention, Hanely doesn't.

The conclusion of these mediation studies and my own inference from a close reading of the empirical studies are essential for the Assessment Framework developed in the last chapter. The empirical study of Abbasi concludes that non-judgmental facet is most effective for fostering the self-knowledge which then foster the health benefits. This means that we can rely on it as the centerpiece of our Assessment Framework. The Assessment Framework uses the opposite of non-judging awareness, namely, judgmental awareness, to assess how much to trust a person's privileged self-knowledge claim in a target peer disagreement. With the empirical

research of Abassis, Hanley, Sala, Hayes, Lattimore, Levoy, Hsu, Verrier, and Morillo-Sarto, we can be much more confident that the use of "judgmental awareness" in the Assessment Framework is reliable.

Another amazing benefit for the Assessment Framework is that we now with this study have a way of scaling the Assessment Framework. Abbasi, Hanley, and many other mindfulness researchers use the FFMQ to scale their own statistical studies about mindfulness. And we will see later how the Assessment Framework can similarly use FFMQ to scale trust in privileged self-knowledge claims.

Chapter Three

Strong Correlation and the Fragility of Privileged Self-Knowledge

In this chapter I argue that when the two methods work together more effectively with mindfulness, we can see both that mindfulness is needed for privileged self-knowledge and that privileged self-knowledge claims are limited. To establish this conclusion, I first show how empirical research proves adequate mindfulness is strongly and positively correlated with privileged self-knowledge. When each method is working together better with mindfulness doing what each uniquely does, privileged self-knowledge can occur. Next, we show how the work we have done vindicates Moran against the claim made by Carruthers that people who believe there is privileged self-knowledge are just wrong. I next show how Moran isn't completely vindicated because he doesn't adequately acknowledge the fragile nature of privileged self-knowledge claims due to the many psychological, physical, and social barriers to self-knowledge, a fragile nature that Carruthers adequately acknowledges. Finally, we show how all of what has been said implies a middle position whereby, on one hand, the extreme view claiming people never have such privileged access (represented by Peter Carruthers) and, on the other hand, the opposing view claiming people normally do have privileged access (represented by Richard Moran). Moran is conditionally right that we do have first-person authority; that is, we have first person authority when in the adequate mindfulness condition since adequate mindfulness has been shown to remove inadvertently the ever-present threats of debilitating psychological, social, and physical barriers to such authority, barriers which Carruthers and empirical studies on self-observation so clearly point out. Privileged self-knowledge claims typically reflect the reality of a person's mental state given adequate mindfulness, yet their fragility makes them susceptible to defeat. This chapter argues for the following statement of this middle position:

The middle position on claims to self-knowledge based on privileged access:

Though a person has privileged access to a belief about the self resulting from deliberative agency even if the propositional content of that belief is false, the propositional content of such a privileged belief is likely true given adequate mindfulness such that we appropriately assume its truth prima facie; but because such claims are so easily susceptible to psychological, social, and physical barriers to privileged access leading to false propositional contents, the prima facie status of such claims is fragile.

We must say a couple of things before starting the process. The focus we place on Moran's work rather than the work of others who affirm privileged access (Bar-On, 2004; Korsgaard, 2009; Shoemaker, 1994; Wright, 2015) is appropriate because the empirical research on mindfulness also affirms many other things about Moran's specific view of privileged access. The middle position makes possible the Prima Facie norm, which in turn makes possible the Assessment Framework.

Adequate mindfulness provides context conducive for privileged self-knowledge

With what we have said so far about the substantive self-knowledge produced by mindfulness in Chapter Two, we can answer the following question: Does privileged self-knowledge require adequate mindfulness?

To see how this question can now be answered we start first by explaining how everything we have said about the conscious-created and substantial self-knowledge produced inadvertently by mindfulness implies that this self-knowledge is privileged. We have seen how the mindfulness-generated self-knowledge has the attributes of a propositional attitude that Carruthers calls "substantial" and that Moran calls first-person authority. This self-knowledge, as we have seen, is final in the sense that it settles the question. For those subjects in the Sala study and the many similar studies, the self-knowledge that replaces the false belief is the final say in how one thinks about the issue. The subjects also decisively commit to the belief. They fully engage the replacement belief. This self-knowledge is actionable in the sense of being "immediately and non-inferentially available to inform practical reasoning" (Carruthers, 2011, 103-4). Subjects in the Sala and similar studies immediately use the self-knowledge to influence their practical behaviors.

The self-knowledge produced inadvertently by mindfulness also has the two qualities that Moran says are necessary for privileged self-knowledge, non-sensorily derived and non-interpretive. When mindfulness engages its enhanced sensory observation, it notices that some things observed function as sensory cues that, when triggered by observation, activate embodied interpretations about what one's mental state is. Here interpretations are pointed out as interpretations, whereas ordinarily they would not be seen as such; one normally just yields without thinking to the cues and their consequent interpretations. At this point mindfulness

can't judge the interpretation, since judgment isn't part of its remit. In just pointing out the cued interpretations as interpretations, one has some reflective distance from the interpretation. Sala and others describe mindfulness as giving subjects a crucial reflective distance from what are normally automated interpretations, and this reflective distance allows the individual to make her own conscious decision about the interpretation. What this means is that mindfulness provides a context of reflective distance in which an individual can consciously evaluate the interpretation. In making a decision about an interpretation that shows up in the enhanced observation of mindfulness, one isn't completely influenced by the interpretations. The self-knowledge known to be produced by mindfulness deliberates without automated interpretations activated, settles an issue in a final way, and is available for practical reasoning; all of these characteristics make this self-knowledge privileged.

The reflective distance produced directly by mindfulness, then, highly facilitates privileged access. Without the distance that allows one to see interpretations as interpretations, one can't make a free and deliberative decision to accept or reject the interpretation. Deliberation requires a conscious awareness which weighs the merits of the interpretation and often weighs them in relation to alternative options. You can't truly decide consciously what your decision on an issue is, and you can't see a decision as <u>your</u> decision consciously made, unless you have reflective distance from the view. Mindfulness fulfills that highly conducive condition.

The conscious decision of mindfulness-produced self-knowledge also is not directly influenced by sensory inputs. In the self-knowledge known to be produced by mindfulness we have been considering, mindfulness, with its enhanced sensory observation, points out when something observed triggers an interpretation. The awareness of the interpretation as an interpretation provides an opportunity for the agent to deliberate whether or not to accept the interpretation. The sensory cue no longer determines one's decision. With the reflective awareness of an interpretation as an interpretation it is possible that a conscious deliberation can take place and decide the issue; the mere observation of a cue no longer decides the mental state which one is in with an interpretation put in place by unconscious deliberation. In this sense the conscious deliberation of the privileged self-knowledge produced by mindfulness is not sensory.

It is important to see that privileged self-knowledge claims are more reliable when observational-interpretive and privileged access methods work together more effectively with mindfulness. The goal of the observational-interpretive method is to accurately observe a cue, to accurately understand what interpretation is embodied in the cue, and to act on that interpretation assuming that it will reliably lead one to believe one has the mental state one actually does have. Because mindfulness has enhanced observation and non-judgmental awareness, it helps one attain the first two of these three goals. Mindfulness can't act on the interpretation embodied in the cue because its sole remit is to observe in the best possible way. The goal of the privileged access method is to deliberate about options for solving an issue and to decide the issue freely without being influenced by ulterior motives, automated interpretations, and biases. Because mindfulness reduces dysphoria with its non-judgmental aspect making higher quality information available and ulterior motives unnecessary, and because mindfulness yields reflective distance from interpretations, the privileged self-knowledge claim is, with high mindfulness, more reliably able to avoid being influenced by automated interpretations and biases that compromise the freedom of its deliberation. Thus, with mindfulness each of the two methods for gaining self-knowledge does what it does more effectively.

Also, privileged self-knowledge claims need more balanced and higher quality information given through mindfulness' enhanced observation in order to avoid bias and automated interpretations. Barriers to privileged self-knowledge are put at bay by the reflective distance that mindfulness affords. And this means that privileged self-knowledge claims are more reliable the more the first two goals of the observational-interpretive method (to accurately observe a cue, to accurately understand what interpretation is embodied in the cue) are effective with mindfulness. Thus, the two methods work together more effectively with mindfulness to attain more reliable privileged self-knowledge claims.

We now have all the components of what Moran calls privileged self-knowledge. The self-knowledge produced by mindfulness originates from conscious deliberation, isn't influenced directly by sensory or interpretive input, decides and settles an issue in a final way, commits to a belief, is used to address practical issues, and isn't derived from anything other than a conscious deliberation. The person with this self-knowledge knows it in a way that others can't. This is so because the self-knowledge is constituted by the act of deliberating

and deciding the issue. Transparency is the quality of a decision or judgment whereby when making the judgement or decision one in real time knows one's own mental state. The propositional attitude is transparent in the act of deciding an issue about the self. We are aware of substantial beliefs about the self exactly the instant they are formed because the substantial belief is formed in the act of deciding and one is conscious of the act of deciding. Nobody other than the one deliberating and deciding an issue can have that immediate act of deciding. Others could possibly know what my mental state is by observing my behaviors or listening to me tell them what it is, but only I can know it directly without interpretation and observation. Therefore, the self-knowledge shown to be produced by mindfulness certainly is privileged.

We can now definitively answer the question: Does privileged self-knowledge require adequate mindfulness? The answer is yes. Adequate mindfulness is strongly and positively correlated with privileged self-knowledge at least for the kind of privileged self-knowledge that is shown to be produced by mindfulness.

But mindfulness isn't the only thing that one must have for there to be privileged self-knowledge. Adequate mindfulness is a highly conducive, but not a sufficient condition. While the sufficient condition of privileged self-knowledge is the conscious and free deliberation of the agent settling an issue, adequate mindfulness highly facilitates the deliberation to be actually determined by the agent freely provides a context where one has reflective distance seeing automated interpretations of one's mental states as interpretations.

Mindfulness isn't just a highly conducive condition for privileged self-knowledge because it allows one to have the reflective distance from one's own standing interpretations such that one can freely deliberate among alternative options, whether or not one ultimately chooses the same interpretation. It also is a highly conducive condition for privileged self-knowledge because it allows one to have the reflective distance from the many psychological, physical, and social barriers that influence one's deliberation without one knowing it. While one needs reflective distance from automated interpretations of one's mental states we talked about earlier, one also needs reflective distance from other related barriers to privileged self-knowledge.

We can understand these barriers better by thinking about them as the result of judgmental awareness. We have said that mindfulness is most frequently defined as non-judgmental awareness. Mindfulness works by seeing judgments as judgments and as interpretations. The barriers to privileged self-knowledge are judgments

that influence one's deliberation without one knowing it such that that deliberation isn't actually conscious, isn't actually decided by the agent who isn't even aware of them. In this way the barriers are the result of judgmental awareness. The following are some instances of judgmental awareness:

Examples of judgmental awareness

1) Clear indications a person is jealous with the person refusing to even consider whether the behavior indicates jealousy.
2) A person who radically insists on not taking observational-interpretive evidence that he is delusional seriously when authoritative and relevant psychological tests indicate such.

The detection of judgmental awareness which mindfulness does is very important for privileged self-knowledge. We will soon see that mindfulness removes judgmental-awareness barriers to privileged self-knowledge in many ways. We will see that higher quality and balanced information about oneself is available with mindfulness's non-judgmental awareness which reduces the dysphoria that tends to make people suppress negative information about oneself. We will see that mindfulness has been proven to reduce the judgmental awareness of racism, implicit bias, automated chauvinism, automated views about inferiority, and many other instances of judgmental awareness. We have already seen how mindfulness can reduce the dysphoria normally present when recognizing an unflattering truth about oneself such that judgmental awareness doesn't need to cover up that truth. We will see studies that show mindfulness' ability to reduce dysphoria such that false judgmental awareness can be replaced by more accurate self-knowledge. Mindfulness provides a context where one's deliberation is free of unaware barriers that get in the way of a free decision. We will see that mindfulness facilitates a context where one acknowledges the many barriers to truly self-deliberated decisions, barriers that can make one think that one is deciding an issue about oneself when in fact it is being decided for one by a psychological, physical, culture influence or ulterior motive. Adequate mindfulness in this way provided a highly conducive context at least for the extremely consequential types of self-knowledge where one replaces a longstanding and consequential false belief about the self.

More on how mindfulness works producing privileged self-knowledge; how Carruthers overgeneralizes

Self-concept

To see better how Carruthers is wrong in his overgeneralization, we must understand the psychological concept of self-concept, since it will be crucial to see how mindfulness inadvertently fosters direct access to the

substantial mental states—beliefs, judgments, decisions—that populate one's self-concept. Psychological research has proven that there is a temporarily stable, yet dynamic, repository of attitudes towards oneself and of skills that one believes one has, and researchers call this the self-concept. We will see that the research on mindfulness concludes that mindfulness inadvertently helps people have clarity about the self-concept. The research clearly shows that people having the mindfulness trait in a higher level are more aware of their self-concept (SC), and they consequently have more of self-concept clarity. The self-concept influences one's judgments and behaviors, and the more self-concept clarity, the more one knows about one's psychological characteristics, and the more one knows about what influences one's judgment, moods, and behavior. Consider the definitions:

Self-Concept

> One's description and evaluation of oneself, including psychological and physical characteristics, qualities, skills, roles and so forth. Self-concepts contribute to the individual's sense of identity over time. The conscious representation of self-concept is dependent in part on nonconscious schematization of the self. Although self-concepts are usually available to some degree to the consciousness, they may be inhibited [emphasis mine] from representation yet still influence judgment, mood, and behavioral patterns. (American Psychological Association Dictionary of Psychology, 2007)

Self-Concept Clarity

> The extent to which the content of an individual's self-concept is clearly and confidently defined, internally consistent, and temporally stable. (Campbell, 1990, 538)

Notice the word "inhibited" in the definition of self-concept. We will claim that mindfulness inadvertently removes the hidden barriers that inhibit self-concept clarity. Self-concept isn't a fixed essence of the self, an unchanging true self, or a soul. Rather, it is a relatively stable—part consciously understood, part unconsciously schematized—identity that influences how one thinks and acts, and that incorporates key judgments and dispositional assessments about oneself and one's relations to the world. A person is certainly aware of some of her relatively stable identity, though aspects of the self-concept are often hidden or inhibited even while those hidden parts still influence one's judgements, moods, and behaviors. Of course, self-concept can change throughout one's life. We are not born with a self-concept.

Inadvertently fosters a channel for privileged access

As we have established, mindfulness inadvertently facilitates what Carruthers thinks isn't possible, the ability to know one's own substantial mental states directly without sensory processes, without observation, and without interpretation, that is, knowledge of oneself based on direct access.

Direct access:
Direct access: non-sensory, non-observational, non-interpreted awareness.

On the face of it, it seems ridiculous to think that something whose only function is to intensely and without judgment observe fosters inadvertently access to what isn't known through observation, sensory input, and interpretation. Yet, empirical research on mindfulness proves that the higher the level of mindfulness, the higher the level of clarity about crucial and consequential mental states of one's self-concept. In the presence of this kind of intense and non-judgmental observation, mindfulness, we likely have awareness of our most consequential mental states. Mindfulness inadvertently fosters a communication channel between the self-concept and one's conscious awareness of oneself.

The studies of David Vago show that people in the mindfulness condition have better self-concept clarity, that more self-concept clarity results in better decisions about the future, and that those better decisions yield more satisfaction and increased happiness (Vago, 2012). Above we showed how the self-concept clarity is known in a privileged way. In a separate empirical study, Adam Hanley in 2017 corroborates exactly the conclusion of Vago in 2012, namely, the more dispositional mindfulness one has the more one has self-concept clarity (Hanley and Garland, 2016, 337). Dispositionally mindful people have substantial self-knowledge when they relax habitual and automated beliefs about the self, and this increases the clarity with which the nature of the self is perceived. Hanley and Garland take it a step farther with their findings, saying, "It appears that dispositional mindfulness may occasion greater clarity with respect to beliefs about the self." It creates less biased beliefs about the self making the self-concept more stable (Hanley and Garland 2016, 337). This means that mindfulness results in decisions about the future that reflect the judgments and attitudes of one's self-concept which is temporary, though stable and dynamic. Hanley says the self-concept clarity fostered by mindfulness provides "a stable platform from which autonomous decisions can be made in reference to an idealized self" (Hanley and Garland 2016, 337). In the mindfulness condition, a decision about the future is

made with self-concept clarity. This means one is conscious of the very consequential judgments and attitudes of one's self-concept; the decision is made on the basis of this awareness. Above we have shown how the self-knowledge results of mindfulness are attained non-observationally. And the fact that the self-knowledge results are attained in a non-observational-interpretive way means that mindfulness only inadvertently produces that self-knowledge, since mindfulness only observes in a non-judgmental way and so can't directly produce something attained without observation. Also, other research shows that those with more self-concept clarity due to mindfulness have less ambiguity about their propositional attitudes (Dummel, 2018).

Even more research supports the already compelling evidence presented by Dummel, Vago, and Hanley that mindfulness indirectly fosters occurrent knowledge of one's attitudes, substantial mental states, and self-concept. For example, a study finds that people higher in mindfulness are less prone to impact bias which overestimates a future event's emotional impact (Emanuel et al. 2010). The ability to predict such things accurately is very important for many of life's decisions (Wilson and Gilbert 2005), like for determining whom to marry or what occupation to pursue. Those higher in mindfulness more accurately predict how they would feel about a future event, for example, whether one will be happy in future situations. Emanuel and colleagues conclude that, "a mindful perspective allows individuals to be more cognizant of how life events influence their emotional experiences, enabling individuals to make predictions that are less susceptible to the impact bias" (Emanuel et al. 2010, 815). They point to a specific quality of dispositional mindfulness that reduces impact bias: Mindfulness's focus on observing one's occurrent and inner emotional experiences in interaction with external events. Of course, mindfulness is observational, but it is the enhanced and non-judgmental observation of mindfulness that helps one see the automated and sensory-based ways that one comes to know oneself by, thus having a distance from such sensory means such that one can know in a privileged way how one's emotional experiences interact with external events. In their own words, what reduces impact bias is, with mindfulness, "a more general knowledge of how both internal and external events affect emotions" (816). For prediction of one's future happiness to be more accurate in mindfulness conditions, one would need to be aware of the substantial mental states deciding how internal and external events influence one's emotions and future

happiness. Also, people with mindfulness training appear to know better their feelings of self-worth (Koole et al. 2009), something that surely meets at least some of Carruthers' criteria for substantial mental states.

Inadvertently disrupting interpretations normally unconscious

While we have proven already that mindfulness inadvertently fosters privileged self-knowledge by disrupting interpretations normally unconscious and automated, we can add more evidence to this conclusion. As we have seen, system-one interpretations are automatically and unconsciously generated in response to cues that we observe in our environment. As we have pointed out above, mindfulness disrupts these automated interpretations by seeing them as interpretations. Normally we just engage automatically interpretations that are triggered by cues in our environment. Because of its enhanced and non-judgmental observation, mindfulness allows us to see those interpretations as interpretations. It seems odd to say that mindfulness "sees" these interpretations triggered by cues, since interpretations themselves are not physical. An interpretation, in the way that we are using the term here, is just a significance (or meaning, purpose things are put to, function, fetish, symbol, value, etc.) of something in the world that goes beyond just the thing itself and influences how people respond to the thing or situation. Yet mindfulness sees the interpretations by their effects in a similar way as astrophysicists see planets outside our solar system by detecting the disturbances in starlight around the planet. Similarly, mindfulness doesn't see the interpretations imbedded in cues directly; rather, it sees the influences and effects of the interpretation.

Interpretations =

An interpretation, in the way that we are using the term here, is just a significance (or meaning, purpose things are put to, function, fetish, symbol, value, etc.) of something in the world that goes beyond just the thing itself and influences how people respond to the thing or situation. For example, a Christian my see two sticks crossing each other as a cross which has significance way beyond just the fact that they are two sticks.

Seeing interpretations as interpretations =

The unobservable significance (meaning, purpose things are put to, function, fetish, symbol, value) of things or situations, seen by the observable impact that it has on observable things, is understood to be one interpretation among many possible ones.

The significance placed on something in the world, seen by the influence it has on the things around it, can be as simple as noticing that the particular wooden thing with metal on the end is a hammer or as complex as noticing that the two long beams crossing each other with nails on the ends is a sign of salvation for Christians. In both of these extreme cases there is a significance placed on them that influences what one can observe, that

is, how they are used, how they are valued, how people act in front of them, the use that people put them to, whether people put them in the garbage after using them, or whether people put them at the front of a huge building and worship them. The significance is something greater than its parts.

The research below indicates that the privileged self-knowledge inadvertently fostered by mindfulness results largely from its ability to see these interpretations as interpretations. Seeing the interpretations as interpretations does a number of things conducive to privileged access. Pointing out interpretations as automated interpretations (that is, a meaning is associated with a thing unconsciously) makes it less likely that we confuse these automated and unconscious interpretations for truly deliberative decisions (that is, decisions that one carefully thinks about consciously) or judgments, which is the mistake Carruthers says we always make when we claim to have consciously deliberated decisions. Seeing an interpretation as an interpretation also provides an occasion to think in a more deliberative way about an issue, concern, or problem. Because mindfulness is known to indirectly foster privileged access, the at least slight disruption of the automated interpretation present when one sees the interpretation as an interpretation likely can lead to truly deliberative decisions and judgments that reflect one's self-concept.

Mindfulness has been proven to do these things through many studies. Sala, as we have seen, proves that mindfulness helps reduce emotional eating by interrupting automated interpretations of cues in subjects' environments. These cues can be internal events such as thoughts or feelings. By seeing the automated interpretations of cues as interpretations, mindfulness interrupts the automation. Once the person is conscious of the interpretation, that person is more likely to think about other options (Sala, 2020). (Lueke and Gibson, 2015) find that mindfulness works to reduce racial and ageist biases by interrupting the automated interpretations triggered by cues in one's environment (Lueke and Gibson 2015, 285). Another similar study concludes that people with food cravings, after a mindfulness exercise, attained a better self-knowledge of cravings and their ephemeral nature that diminished reactivity to them (Papies, Barsalou, and Custers 2012). They know better what the cues and interpretations are of their addiction. As Papies and colleagues state, "mindful attention prevents mindless impulse" (Papies, Barsalou, and Custers 2012). (Kernis and Goldman, 2006) says with mindfulness there is increased knowledge of self-motives.

Mindfulness doesn't tell the addict to stop eating sweets, stop using cocaine, stop hitting a spouse in a flurry of rage, stop ruminating, or love yourself. These judgments aren't part of mindfulness's remit. Any of the health and self-knowledge benefits of mindfulness have to be the inadvertent byproduct benefits of its passive remit. When people see the interpretations as an interpretation without judgment just as one among many possible interpretations, this often frees people from seeing interpretations as absolute and unchangeable; that inflexible way of thinking often causes addiction and dysphoria, as we will see. A ground swelling plethora of empirical research confirms both that the health and self-knowledge benefits of mindfulness work by calling attention to the interpretations generated in the observational-interpretive access method for knowing oneself, and confirms that this often inadvertently makes possible the disruption of the observational-interpretive process.

Helen Ma and John Teasdale conclude people prone to depression often have a tendency when they experience negative, dysphoric thoughts to ruminate on them, and rumination leads to relapse (Ma and Teasdale 2004). It isn't necessarily the negative, dysphoric thoughts that lead to relapse, rather it is the way one reacts to and interprets them with impacted ruminations that exacerbates the problem. They often chronically move automatically from the dysphoric thoughts and feelings to rumination about them, and mindfulness helps depressed people stop this automated transition. Depressed people seeing the disconnect through mindfulness are consequently freed up to consider different ways of relating to the dysphoric thoughts that can be more helpful. Mindfulness disrupts debilitating cognitive sets and interpretations, and it helps one replace them with more adaptive strategies.

Another study finds that individuals who score higher on dispositional mindfulness are better at differentiating their discrete emotional experiences with less emotional difficulties and more effective emotional regulation (Hill and Updegraff 2012); this study didn't engage any special mindfulness training, rather, researchers measured the variety of dispositional mindfulness and its effects.

Inadvertently maintains the channel by blocking reactive barriers

In what follows we will show how mindfulness does many other things to inadvertently foster a channel between one's self-concept and one's conscious deliberations. It blocks the unconscious motivations that

inhibit self-knowledge. There is much evidence that mindfulness inadvertently disarms debilitating unconscious motivational barriers making accessible a higher quality of information about oneself. It does this by inhibiting the reactivity to dysphoric mental states (e.g., ego threatening information) that activates the ulterior motives causing confabulated interpretations, ulterior motives which have goals other than attaining the truth of the matter about one's substantial mental states, like excessively making oneself look good or hiding information about oneself that isn't pleasant. We will soon see the empirical evidence for this, but for now let's look at how this inhibition is thought to happen. When inhibiting the reactivity and dysphoria, one inhibits as well the ulterior motives that kick in to help one deal with the uncomfortable reactivity. We are calling "ulterior motives" motives not aimed at attaining the truth of the matter and often kept hidden in order to get a desired result. Ulterior motives are barriers to reliable self-knowledge since they aren't aimed at attaining the truth about one's mental states. Erika Carlson, a foremost empirical researcher on mindfulness and self-knowledge, describes motivational barriers as "instances when ego-protective motives influence the way people process and utilize information about their personality" (Carlson 2013, 175). As barriers to self-knowledge, the ulterior motives influence negatively the quality of information the subject has access to with their drive to protect the ego at all costs. The two most dominant motivations are self-enhancement and self-valuation (Sedikides 1993) (Sedikides 2007). As mentioned people have a bias when taking in feedback from others which favors positive over negative information, and the biases are due to "the motive to maintain and elevate the positivity of the self-concept" (Sedikides 2007, 1), which is called the self-enhancement motive. The work of Sedikides and others clearly establishes that the ulterior motives produce lower quality information about oneself (Sedikides 1993) (Alicke 2012). Another bias which favors positive over negative information about the self is "the motive to protect the positivity of the self-concept against threatening information" (Sedikides 2007, 1), and this is called the self-protection motive. There are also truth-conducive motives such as the self-assessment motive. Motivational barriers hide the truth of the matter regarding one's mental states so as to avoid acknowledging underlying feelings, thoughts, and other mental states which go against the projected self-concept. As barriers to self-knowledge, the ulterior motives influence negatively the quality of information the subject has access to. Reactivity causes increased discomfort, and the ulterior motives relieve this discomfort by selectively taking

in data favoring that information that preserves self-image, and this is how ulterior motive activation results in lower quality information. Mindfulness facilitates more reliable self-attributions by reducing the source of troubling reactivity that often unconsciously motivates people to submerge mental states replacing them with confabulated interpretations. Mindfulness significantly inhibits the need for ulterior motives with its reactivity-reducing non-judgmental way of observing. Mindfulness isn't imposing an interpretation on the reactivity, rather its non-judgmental remit inadvertently reduces the need for unconscious ulterior motives.

The evidence that mindfulness inhibits the reactivity to dysphoric mental states thus producing higher quality information about oneself is overwhelming. One study examines the memory of depressed people and formerly depressed people with past suicidal tendencies. People with mindfulness training remember with enhanced detail dysphoric memories about their previous episodes of depression, and the particular non-judgmental way of encountering such details helps people avoid memory inhibiting reactivity (Hargus et al. 2010). Other researchers, for example, John Teasdale and Williams, earlier found similar results (Williams et al. 2000). Mindfulness facilitates higher quality information about the self, and that means more balanced information is available where positive information about oneself isn't favored over negative. Along with Helen Ma, Teasdale recreates his earlier study just mentioned of 2000 and confirms the earlier conclusion that mindfulness is very effective for helping people prone to depression avoid relapse through its effectiveness at reducing reactivity (Ma and Teasdale 2004). Also, people trained in MBSR (Mindfulness Based Stress Reduction) and shown a sad film demonstrate less rumination and neurological reactivity to their dysphoria as shown in fMRI imagining even though they report the same amount of such dysphoria as those untrained in MBSR (Farb et al. 2010, 31). The authors conclude people prone to depression often have a tendency when they experience negative, dysphoric thoughts to ruminate reactively about them, and rumination leads to relapse. It isn't the negative, dysphoric thoughts that lead to relapse, rather it is the interpretive way one reacts to them that leads to impacted and debilitating ruminations. Similarly, (Arch and Craske 2006) finds that people with mindfulness take in more balanced information when presented with negative slides due to less reactivity. (Creswell et al. 2007) and (Way et al. 2010) produce neurological evidence for a decrease in negative affect and less reactivity; reactivity clearly produces lower quality information. Mindfulness has been shown in these

and other studies to reduce amygdala activity, and higher amygdala activity is a sign of more reactivity. With mindfulness' reduction of reactivity more balanced information can be processed in a way that causes less dysphoria and activates the prefrontal cortex while decreasing the amygdala. There is more balanced information coming in because of the inhibition of reactivity, and this helps people avoid depression. (Modinos, Ormel, and Aleman 2010, 369) demonstrates that mindfulness reduces reactivity, and this helps people control negative emotions: "These findings suggest that individual differences in dispositional mindfulness, which reflect the tendency to recognize and regulate current states, may modulate activity in neural systems involved in the effective cognitive control of negative emotion." Notice that this and other studies find and analyze naturally developed mindfulness in individuals, and so we are not just talking about correlations due just to intense training. Indeed, Ruchika Prakash and colleagues find that dispositional mindfulness avoids the maladaptive reactivity resulting from thought suppression and the rebound that it leads to (Prakash et al. 2017, 84). Another study finds that individuals who score higher on dispositional mindfulness are better at differentiating their discrete emotional experiences with less emotional difficulties and more effective emotional regulation (Hill and Updegraff 2012); this study didn't engage any special mindfulness training, rather, researchers measure the variety of dispositional mindfulness and its effects. In another study, researchers find that people actively engaging mindfulness exercises report increased awareness of repetitive thoughts during the exercise and are less emotionally reactive to their repetitive thoughts compared to two other stress management therapies, progressive muscle relaxation and loving-kindness meditation (Feldman, Greeson, and Senville 2010, 1008). Again, people with higher mindfulness report less reactive anxiety in a social stress test with lower cortisol responses (Brown, Weinstein, and Creswell 2012); (Britton et al. 2012) finds something similar. Certainly, mindfulness facilitates more access to quality information about one's mental states because of reduced reactivity. The presence of higher quality information in the mindfulness condition, just illustrated and substantiated more in the next section, is evidence that there are fewer confabulation barriers. The confabulated interpretations Carruthers and Wilson talk about hide the truth about one's own mental states by placing guessed or fabricated self-understandings in the way of any potential higher quality information

revealing the truth about the self. With mindfulness' barrier inhibition, there is inadvertently more direct access to one's own mental states.

Mindfulness studies in relation to hallucination support the view that mindfulness indirectly fosters self-knowledge and health benefits by reducing reactivity. In studies of the impact of mindfulness on people prone to hallucination, researchers find that while those with mindfulness show an enhanced awareness of occurrent negative thoughts and feelings generated at the present time, this enhanced awareness is less threatening because of the non-judgmental nature of mindfulness. Proneness to hallucination is positively correlated with ruminative self-focused attention, and mindfulness decreases it with its non-judgmental observation (Perona-Garcelán et al. 2014, 6); mindfulness allows "negative thoughts to go by without reacting to them" (6). People with higher mindfulness tend to think negative dysphoric thoughts aren't necessarily representative of reality and aren't necessarily representative of one's person (Perona-Garcelán et al. 2014), and surely that is a substantial mental state made accessible inadvertently through mindfulness. Debilitating reifying interpretations of dysphoric thoughts are inhibited in the mindfulness condition. The act of seeing thoughts as not necessarily representative of one's person is called decentering in the literature, and (Lueke and Gibson 2015) gives an excellent list of studies supporting the view that mindfulness works by decentering the self and by reduction of reactivity.

Moran's deliberative agency conditionally supported by mindfulness

The conclusion that we have reached has implications for Moran's work. Mindfulness does one thing very helpful for Moran. Moran lumps all the things that foil first-person authority, making it such that the propositional content of beliefs showing up in rational deliberation are false, into the category of "special cases." The research on mindfulness gives the Moranean view a clear picture of what those special cases are. They have to do with the barriers to authentic first-person authority. Mindfulness studies show us what has to be done in order to assure these special cases don't happen. Moran doesn't spell out these things. And we need to, since we can't attempt to avoid special cases if we don't know how to do this.

Mindfulness research supports many essential aspects of Moran's views. The most important aspect of Moran's work is that first-person authority comes naturally and transparently in the process of rational deliberation solving an issue. There is no active mediation of an inner sense. And mindfulness studies do this without positing an inner sense or some other mechanism that produces directly the privileged access. Like Moran's view, self-knowledge produced in the mindfulness state isn't produced by inference, evidence, or observation; the authority just naturally shows up in the process of consciously deciding an issue.

With help from mindfulness studies, we declare victory for Moran over Carruthers in the issue of whether or not we have privileged access to our substantial mental states. But Moran's victory must be reframed in light of three things: 1) what Carruthers' work has correctly pointed out about the importance of observational-interpretive access, 2) what empirical studies on self-observation show us about how interpretive-sensory processes often go wrong even though they often go right, and 3) the discovery that mindfulness serves to inadvertently manage observational-interpretive processes so that genuine privileged self-knowledge is more likely. By "manage" here we are referring to the two things that we have seen mindfulness inadvertently do, blocking the need for barriers and disrupting automated interpretive sensory processes, not by necessarily ending them, but rather by seeing their interpretations as interpretations. It's a passive management. It merely inadvertently keeps everything on the right track for fostering privileged access, even though it doesn't intend to do this. To get to a point where we can see the views of Moran that mindfulness studies corroborate, on one hand, and find wanting, on the other, we must first understand Moran's realism.

Moran's realism

I will argue that Moran shows how Carruthers' conditions listed for a truly deliberative decision are satisfied often. In order to get to a point where this makes sense, we will first need to describe Moran's "ordinary realism."

It is important to see that Moran is a realist about the things we know about. Moran in no way thinks of his work as going against what he calls "ordinary realism." Early on in his major book, when he gives a short

description of what is in the different chapters of his book, he describes ordinary "realism" in the following way:

> Ordinary "realism" about the mental suggests a relation of logical independence between the description of some feature of mental life (e.g., a thought or emotional response) and the feature or state itself. (Moran, 2001, xiv, see also 37)

Realism about the mental implies a separation between the description of mental life and the state of mental life itself. We will see soon exactly what this separation involves.

> We should not confuse the introduction of the agent in self-reflection with either abandoning ordinary realism about the mental or denying a substantial epistemology for self-knowledge. (Moran, 2001, 59)

His concentration on the exclusive access one has due to deliberative agency doesn't deny ordinary realism about the mental state. He wants to make sure that people get it right about the relation between the psychological facts observable by anyone as the objects of realism ("the state itself") and the first-person deliberative agency exclusively known only for the specific agent ("the description of some feature of mental life").

In order to express this relation more clearly, Moran points to situations where the two go apart, that is, empirical psychological facts and the facts of practical-deliberative agency. Moran talks about how one can have a "sophisticated vocabulary for self-interpretation" and at the same time have an illusion about the actual state of mind (Moran, 2001, 42, 50). Consider the passage below where Moran essentially says one can have privileged access to one's belief about one's own mental state even when the propositional content of the belief is false:

> There are a number of ways in which a person's reconception of his state of mind will require an altered description of his state, and for reasons that are first-personal, not shared by anyone else's conception of him. There is a sense in which such an idea accords a special "privilege" to the person's self-conception, since it is only his own conception of his state, and no one else's, that is claimed to have this logical character. But the idea of "privilege" here should not prevent us from seeing that this status given to the person's own conception does not depend on his interpretation being *true*, let alone true because it is self-constituting. One reason for this is simply that even someone's false conception of his state is part of the very person we want to understand ("what must his envy really be like if he's inclined to misdescribe it in this way?"). Even someone's fairly gross misrecognition of his desire or fear will nonetheless be an important indication of the nature of his attitude itself.… Admitting all this, however, need not prevent us from imagining the case as one in which the sophisticate is seriously wrong, or misguided about himself. There is still room for the idea of accuracy and truthfulness in this domain, and for the attendant risks of error and illusion. His interpretation of his gratitude as resentful does not constitute it as such, any more than the naive person's self-understanding makes it the case that his gratitude is innocent. … Retaining the possibility of being wrong does not mean that we abandon the appearance of a self-other asymmetry here. A false conception of one's state can constitute a difference in its total character, and still be false for all that. Someone may see his pride as sinful, but if there is no such thing as sin (really), then surely his conceiving of his pride this way cannot constitute it as such. … Hence, contrary to what is usually assumed, the hermeneutic privileging of self-interpretations (whether individual or social) does not require the assumption of their truth. (Moran, 2001, 48-50)

The way people feel and think in their deliberative agency can be seriously wrong from the objective perspective of realism. He puts quotation marks around "privileged" in the quote above in order to indicate that there is a sense of "privileged"—albeit this sense of "privileged" is a sense that he doesn't want to focus on—in which what the person is "privileged" to, a belief about oneself for example, can have propositional content that is false. A false conception of oneself is a conception that nobody else has, and so it is "privileged" in this sense. It isn't that he doesn't endorse this sense of "privileged." He recognizes that it occurs in what he calls "special cases," as we will see. He wants to focus on beliefs people understand about themselves whose propositional content is true, since they are the kinds of beliefs that normatively occur. Not enough has been done describing these normative senses of first-person authority and privilege. Leave the "special cases" for others to focus on. He also wants to say that even when the propositional content of one's belief about oneself is false, still that self-understanding influences the objective understanding of the kind of mental state the person has. We can now say clearly what the realist relation is between the psychological facts observable by anyone and the first-person deliberative agency privileged only for the specific agent: we have privileged access to our beliefs resulting from rational deliberations even when the propositional content of those beliefs is false.

A middle position about privileged access

We now have everything we need in order to express the position of this paper between the two extremes. Here is that concluding position once again:

> **The concluding position of this paper between the two extremes:**
>
> Though a person has privileged access to a belief about the self resulting from deliberative agency even if the propositional content of that belief is false, the propositional content of such a privileged belief is likely true given adequate mindfulness such that we appropriately assume its truth prima facie; but because such claims are so easily susceptible to psychological, social, and physical barriers to privileged access leading to false propositional contents, the prima facie status of such claims is fragile.

Think of a spectrum of positions on privileged access say with Moran on the right representing those who think we often have privileged access. There is a sense in which one has privileged access to beliefs about the self resulting from deliberative agency even if the propositional content of that belief is false, and he talks about this situation (Moran, 2001, 48-50). Here the person has made a grave mistake about her mental states. Yet at the same time that belief about the self whose propositional content is false nonetheless is "privileged;" the

person has access to that dubious belief about herself in a way others don't. Moran puts quotation marks around this sense of "privileged" (Moran, 2001, 48-50) because he doesn't want to focus on them, what he calls in the following quote "special cases."

> Suffice it to say that they are taken to have a good prima facie claim to truth which may be overruled only in special cases. The important point is that these are taken to be genuine judgments, expressive of knowledge, which are made without reliance on "external" observation. (Moran, 2001, 10)

He wants to focus on "genuine judgments, expressions of knowledge." Normally one has privileged access to one's belief resulting from deliberative agency with a true propositional content, and for Moran this means that those beliefs about the self have a good prima facie claim to truth. Mindfulness confirms that Moran is right about the prima facie claim to truth. So, we are quite a bit to the right on the spectrum of positions on privileged access.

But, we are pulled back towards the middle position of the spectrum because we find prima facie status to be much more fragile than Moran's "special cases" implies, given all the psychological, cultural, and physical ways in which one can be wrong about the propositional content of one's belief about the self. Carruthers is wrong to think that one never has privileged access to one's mental states. But, because Carruthers does such a good job at describing cases where we think we have privileged access but we don't, he helps us tremendously to have a better sense of how prima facie status is fragile.

PART TWO

The Literature from Inscrutable Segregation to Scrutable Interrelatedness

In Part Two I conclude that the literature on peer disagreement largely doesn't adequately acknowledge the fragile nature of privileged self-knowledge claims due to the many psychological, physical, and social barriers to self-knowledge; and I argue with urgency a path forward by recognizing the interrelatedness of the observational-interpretive and privileged access methods of self-knowledge, since the two methods integrated with mindfulness help disputants acknowledge the fragile nature of privileged self-knowledge claims. To reach these conclusions, I investigate the extent to which the current scholarly literature on peer disagreement acknowledges the fragile status of privileged self-knowledge claims, and I show how five key scholars of peer disagreement rely on key privileged self-knowledge claims that either lead to mistaken assessments of the epistemic status of disputants or lead to the likelihood of such mistakes. To remedy the tendency in the literature to segregate the two methods of self-knowledge thinking the privileged self-knowledge claim is inscrutable, I pull from earlier chapters the evidence that they are interrelated and the privileged self-knowledge claim is indirectly scrutable. The integration of the two methods with mindfulness will lead to more accurate assessments of epistemic status in target disagreements.

What is even more disconcerting about the state of the literature on peer disagreement than mistakes made about the epistemic status of disputants is that the key scholars of peer disagreement promote views that hinder the detection and remedy of the mistakes that we will point out in the next two chapters. In Chapters Four and Five we see how key scholars incorporate privileged self-knowledge claims as crucial parts of their view of peer disagreements, and in the process of this incorporation they inappropriately segregate the observational-interpretive method and the privileged method for knowing one's mental states. This segregation results in what I call the Inscrutability Thesis which says, in abbreviated form (see Chapter Four), because one person can't scrutinize the privileged self-knowledge claims of others in a peer disagreement, it makes sense to favor one's own view over the opponent's when everything else seems equal. This view segregates the two ways of

knowing one's mental states when we have seen in Part One that they are integrated. David Christensen and Michael Bergmann explicitly express the Inscrutability Thesis whereas Jennifer Lackey, Ernest Sosa, and Thomas Kelly more implicitly lean on it. In the coming chapters you will see how these scholars don't recognize relevant mistake possibilities in their epistemic assessment method just because they think one should favor one's own privileged self-knowledge claim since others can't scrutinize it and since you can't scrutinize theirs. The mistakes the key scholars make in the following chapters are made because they don't see how the privileged access method for knowing one's mental states functions more effectively when it is engaged together with the observational-interpretive method informing it when there are observations that don't harmonize with the privileged self-knowledge claim. In other words, the two methods work more effectively when they are integrated with mindfulness.

The two chapters of this Part Two don't just diagnose mistakes other scholars have made and what views lead to the mistakes. They present a remedy for the mistakes and the views leading to the mistakes. I claim that this Inscrutability Thesis is false in Chapters Four and Five, and I show the remedy for this mistake is the Indirect Scrutability Norm which works just because both methods are used together to effectively assess privileged self-knowledge claims. The Indirect Scrutability Norm actually goes beyond integration by acknowledging the interrelatedness of each method, interrelated in the sense that something about one method influences the other. The Indirect Scrutability Norm relies on a deep interrelatedness between the two methods such that the observational-interpretive method can appropriately, but indirectly, scrutinize the privileged self-knowledge method. While more will be said about this later, the interrelatedness means that the observation method can point out behaviors or speech of the claimant that conflict with the claimant's privileged self-knowledge claim, and this counts as evidence that the claimant doesn't have the mental state she claims. The work here developing the Indirect Scrutability Norm crucially helps us derive the Assessment Framework in Chapter Six. The Assessment Framework represents the epitome of the integration of the observational-interpretive and privileged methods for knowing one's mental states.

Chapter Four

Peer Disagreement Literature has not Largely Factored in Fragility to its Peril

This Chapter argues that research literature about peer disagreement has largely not factored in the fragile nature of sincere claims to self-knowledge based on privileged access to its peril, hindering a deeper understanding of peer disagreement. I defend this thesis by pointing out the fragile nature of five privileged self-knowledge claims (each described in its own section) that play key roles in the arguments prominent scholars in the literature make for their views about peer disagreement while inadequately aware of their fragile nature due to the many ways that psychological, social, and physical barriers can prevent people from having privileged claims whose propositional content is true. I find in each of the five privileged self-knowledge claims a key scholar has made a mistake or is highly vulnerable to one just because the fragile nature of such claims isn't adequately acknowledged. And the mistakes and vulnerabilities influence negatively the reliability of any assessment of the epistemic status of disputants needed to figure out who is more likely to get the issue right.

This chapter also argues that there is a common handicap that each of the five scholars has. They don't adequately recognize the ways in which the two methods for self-knowledge can work together with mindfulness to produce more reliable assessments of the comparative epistemic status of each disputant for the purposes of judging whose position is more likely to get it right. In Chapter Three I discuss how the two methods work together with mindfulness to make more reliable privileged-self-knowledge claims. They don't recognize the ways in which the two are integrated. Instead, Christensen's and Bergmann's Inscrutability Thesis, a view that says someone else's privileged self-knowledge claim can't be scrutinized, denies the way in which the two methods are integrated. In fact, it segregates the two methods. While Christensen and Bergmann explicitly use the Thesis, the other three key scholars have this same tendency

implicit in their views of peer disagreement. We derive the Indirect Scrutability Norm as a corrective of the Inscrutability Thesis.

The Indirect Scrutability Norm actually goes beyond integration by acknowledging the interrelatedness of each method, interrelated in the sense that something about one method appropriately influences what the other method does. The Indirect Scrutability Norm relies on a deep interrelatedness between the two methods such that the observational-interpretive method can appropriately, but indirectly, scrutinize the privileged self-knowledge method. I will say much more about the Indirect Scrutability Norm in each of the chapters of this Part Two.

You will see that during the treatment of these five privileged claims there is a crucially important discussion of central issues in the literature in terms of which prominent scholars define their own views without adequately factoring in fragility of their privileged claims. The discussion of central issues aims to show how privileged self-knowledge claims play key roles in those central issues. Only in the process of observing how privileged claims are widely deployed uncritically aware of their fragility by diverse and influential scholars responding to a wide variety of central issues in the field can we speak legitimately about the literature as a whole concluding in the last section that the literature largely doesn't factor in the fragile nature of privileged claims to its peril. The prominent scholars whose thought we describe in this discussion of key issues are: David Christensen, Ernest Sosa, Michael Bergmann, Thomas Kelly, and Jennifer Lackey. For each of the central issues described, we will see that privileged self-knowledge claims play key roles. Here are the key issues that we discuss with a more thorough description below: The Independence Thesis (which determines comparative epistemic status in terms of evidence independent of the initial reasoning), the issue of first-order versus higher-order evidence (When to use appropriately each type of evidence?), the issue of evidence hidden but at the same time effective (When is it appropriate to use hidden evidence?), the Total Evidence view (which says don't slight either type of evidence.), and the Right Reasons view (Evidence of two disputants cancel each other out making first-order evidence decisive.).

We will uncover below the variety of crucial ways in which prominent scholars use privileged claims to define their own views and to respond to central issues in the literature while oblivious to the fragile status of

privileged claims. These are not quotes but rather are paraphrases of the privileged self-knowledge claims they assume:

Used as higher-order evidence confirming first-order evidence (Christensen):
1) "I am highly confident that I am paying very careful attention." (Christensen, 2011, 8-9)

Used to declare privileged self-knowledge of a particular mental state (Ernest Sosa):
2) "I know that I have a headache".

Used as higher-order personal information confirming first-order evidence (Jennifer Lackey):
3) "I have decisive personal information."

Used as primary evidence for a belief (Michael Bergmann)
4) "I am extremely confident that p based on 'I am extremely confident in my insight about the way things really are which implies p.'"

Assigns a basing reason to make rational deliberation possible (Thomas Kelly)
5) "I know my base reasons for my belief, and I know they entail the conclusion."

Privileged claims used to solve key issues

We start the analysis of the five privileged claims by introducing the discussion of key issues that will eventually help us gauge the relevance of privileged claims and their fragility to the literature as a whole. The discussion of key issues describes the fragility of the five privileged claims in the context of some of the most important issues in the literature in terms of which a wide diversity of prominent scholars in the field largely define their own views without taking into account the fragility. And the purpose, again, is to help us assess in the last section the significance of such fragility oversight for the literature as a whole.

We start this discussion with an overview of the spectrum of views in the literature on the epistemic status of individuals in peer disagreements. There are two poles to the issues of whether peer disagreement reduces one's epistemic status. One is the conciliationists—like Richard Feldman, David Christensen, and Adam Elga—who generally think one should reduce one's credence in a belief when engaged in a peer disagreement. At the other end of the spectrum are the steadfastists—like Michael Bergmann, Ernest Sosa, Thomas Kelly, and Jennifer Lackey—who generally don't think reduction is necessarily needed when so engaged, or they think some reduction may be needed but not to the extent of splitting the difference thinking both are about as likely to get the issue right. Steadfastists are more conservative than conciliationists with respect to the recommendation of reduction.

It makes sense that we begin this discussion of key issues focusing on the Independence Thesis developed by David Christensen, since this has been a central issue in the literature (arguably the most central issue) and since many scholars have defined their own views significantly in terms of it. As well as describing the

initiation and development of the Independence Thesis at the hands of the person who champions it, we will describe the most challenging criticisms of it brought by four steadfastists, Ernest Sosa, Jennifer Lackey, Michael Bergmann, and Thomas Kelly.

Here is Christensen's formulation of the Independence Thesis that most everyone cites:

> Independence: In evaluating the epistemic credentials of another person's belief about P, to determine how (if at all) to modify one's own belief about P, one should do so in a way that is independent of the reasoning behind one's own initial belief about P. (Christensen, 2009. 758)

While he has modified that formulation quite a bit over the years since 2009, he has stuck to the core thought here that in a disagreement where one evaluates the epistemic credentials of the other disputant's and one's own reasoning as to whether p, one should determine how (if at all) to modify one's own credence about p by engaging an assessment that is independent of the reasoning behind one's own initial reasoning about p.

We can give here a little background for what Christensen means by "initial thinking." When you are making a calculation just in your head, for example, of 24 times 88, all the thinking that goes into making that calculation is the initial reasoning. Christensen and others call this initial thinking first-order thinking, and when that first-order thinking is used as evidence for a belief, it is called first-order evidence. Christensen also talks about higher-order evidence. Higher-order evidence is information about how reliable your first-order evidence likely is for supporting the target belief. So, for example, if you find that you gave your complete attention to the calculation of 24 times 88, that would be higher-order evidence that your first-order evidence is likely reliable.

Now we can get back to what Christiansen is telling us in the quote just above about how the Independence Thesis works. In that quote he says one should determine how (if at all) to modify one's own credence about p by engaging an assessment that is independent of the reasoning behind one's own initial reasoning about p. He calls this a "reliability assessment" (Christensen, 2018). In this self-assessment, one takes into account all the higher-order evidence one has for the reliability of one's thinking for obtaining true beliefs, and from this viewpoint one assesses whether one should modify the confidence in one's belief. What he means by "modify one's own belief" is deciding whether to give up one's belief or hold it with less confidence.

Let's say for example that you have a disagreement with a friend you think is just as good at mental math as you over the answer to the math problem 24 times 88, and a disagreement ensues with your friend because she comes up with a different answer than you. Both have first-order evidence for their respective different calculation, which evidence is the respective initial thinking when making the calculation. Now you want to figure out whether you should modify your belief as a result of the disagreement with a peer. To decide whether one should reduce one's confidence or stop believing, you must disregard the first-order evidence. Once you bracket out the initial first-order evidence, you must rely on higher-order evidence to assess who is more likely to get the calculation right. If you recall that you gave your full attention to the mental calculation, you then have higher-order evidence that supports a conclusion that you are more likely to get the calculation right. After considering all the higher-order evidence of both disputants in the reliability assessment, you estimate whose belief is more reliable for getting the calculation right, and you adjust your level of confidence accordingly.

Throughout all his corpus, Christensen has said (for example, Christensen, 2007 and 2018) a major motivation for this Independence Thesis is to avoid question-begging in a disagreement. We don't want a situation where each disputant just keeps asserting the truth of a conclusion as a premise to prove the conclusion and to prove that the other isn't an epistemic peer. Suppose two people are having a disagreement about p; one says p and the other ~p. You ask the other why do you believe p, and she just keeps saying over and over again, "p just is the case." We don't want a situation where there now is a disagreement about each person's initial reasoning about p, and each just keeps asserting the truth of their reasoning about p as a premise supporting their own initial reasoning about p; that would be epistemic chauvinism, and there would be no progress in the search for the truth of the matter, or no opportunity for resolving the disagreement in a deeper way.

The way the Independence Thesis is used is such that if, after bracketing out the initial reasoning that p in the disagreement, there is no higher-order evidence that one's own first-order evidence is more likely to be true, then one is obligated to reduce credence; we can see this implication just after his initial formulation of the Thesis in 2009 (Christensen, 2009, 758-9). When you can't give better higher-order reasoning than your initial reasoning that p is more reliable than the higher-order reasoning your opponent gives for her initial reasoning that ~p, that is when you are obligated to reduce credence that p. The accumulation of the higher-order evidence

doesn't just start after the initial thinking has taken place. Rather, one can gather higher-order evidence during the initial reasoning; for example, while making the mental math computation, I could find that I am paying very good attention in the calculation, and such a finding would be higher-order evidence. In fact, Christensen thinks one can even gather higher-order evidence just before the initial reasoning with an estimation of, for example, how much attention one would likely have in the specific context (Christensen, 2011, 10).

It is clear here that Christensen is using an internalist view of justification in his reliability assessment, since the goal of this assessment is to give better reasons that one is aware of for thinking one's own initial reasoning, if one wants to keep one's epistemic status, is better than the other's. Externalists about justification reject this awareness requirement. This makes the Independence Thesis more prone to liberally prescribing reduced credence because it requires people to have better reasons that they are aware of, and it is well recognized that the internalist criterion of better reasons one is aware of is harder to fulfill than externalist criteria for justification, for example the externalist criterion that a belief reliably formed is adequate justification even without needing to be aware of why it is likely reliable.

"I know that I am paying attention" is central higher-order evidence for Christensen

We can now point out one privileged claim needed for Christensen's reliability-assessment whose fragility isn't recognized adequately. This is not a quote but rather a paraphrase of the privileged self-knowledge claim he assumes:

> **Used as higher-order evidence confirming first-order evidence (Christensen):**
> 1) "I am highly confident that I am paying very careful attention." (Christensen, 2011, 8-9)

Christensen uses this privileged claim in the Careful Checking case below (Christensen, 2011, 8-9) as a fact about one's reasoning (FAR) as part of his higher-order reliability assessment that may justify one to stay put in one's credence level when disagreeing with a peer.

> **Careful Checking:**
> I consider my friend my peer on matters of simple math. She and I are in a restaurant, figuring our shares of the bill plus 20% tip, rounded up to the nearest dollar. The total on the bill is clearly visible in un-ambiguous numbers. Instead of doing the math once in my head, I take out a pencil and paper and carefully go through the problem. I then carefully check my answer, and it checks out. I then take out my well-tested calculator and redo the problem and check the result in a few different ways. As I do all of this, I feel fully clear and alert. Each time I do the problem, I get the exact same answer, $43, and each time I check this answer, it checks out correctly. Since the math problem is so easy, and I've calculated and checked my answer so

> carefully in several independent ways, I now have an extremely high degree of rational confidence that our shares are $43. Then something very strange happens. My friend announces that she got $45! (Christensen, 2011, 8)

What is interesting about this Careful Checking case is how Christensen says it is appropriate for one to stay steadfast even when he gives more details about the disagreement after he formally states it initially as it is just above. In a quote below you will see that he endorses ("to a large degree"; Christensen, 2011, 9) the intuition that one should not reduce confidence. Here are the higher-order FARs about the friend and the protagonist in the Careful Checking case, resulting from his reliability assessment, that lead him "to a large degree" to an endorsement of the intuition that the protagonist should not reduce confidence:

> I could see my friend writing numbers on paper and pushing calculator buttons, and that my friend assures me that she did her calculations slowly and carefully, felt clear while doing them, and got her same answer repeatedly. (Christensen, 2011, 8)

> Here, many people feel that I should not reduce my confidence in $43 very far at all. …This intuition — which to a large degree I share — seems to cut directly against Conciliationism, and particularly against Independence. (Christensen, 2011, 8)

> Two people, both generally competent at elementary math, who worked on the same problem, each having done the calculations repeatedly and carefully both on paper and with a well-tested calculator, each having checked the answer in multiple independent ways, each feeling very clear-headed and alert throughout, and each repeatedly coming up with (and verifying) a different answer. (Christensen, 2011, 9)

> While I can definitively rule out the possibility that I've deliberately announced an incorrect answer for recreational, experimental, or performance-artistic reasons, I cannot be nearly so sure of ruling out these possibilities for my friend. Similarly, while I can be very sure that I was actually paying attention rather than going through the motions of checking my answer, I cannot be nearly so sure that my friend was. And while there are conceivable sorts of mental malfunction that would affect my reasoning without my having any sign of trouble, most reason-distorting mental malfunctions come with clear indications of possible trouble: dizziness, seeing patterns moving on the wall, memories of recent drug-taking or of psychotic episodes. And I'm in a much better position to rule these out for myself than I am for my friend. Let me put the information I'm depending on in all these cases under the common label, taken from Lackey, of "personal information". (Christensen, 2011, 9)

The protagonist knows some higher-order FARs about his friend from observation and past experiences with her: The friend also repeatedly used all the different methods—like doing the calculation on a calculator, on paper, and redoing the calculation—to verify her answer always getting the same answer. The friend assures the protagonist that she did her calculations slowly and carefully. Both generally are competent at elementary math.

And he knows through higher-order evidence quite a bit about his own privileged attention claim and doesn't know these things about the other's privileged self-knowledge claim: For example, he knows he can rule out that he is insincere, he knows that he was actually paying attention, and he knows that he was not suffering from a mental problem. The protagonist can be sure that he was actually paying attention. The protagonist knows that he isn't having mental problems or on drugs.

Christensen concludes to a large degree that the protagonist should not reduce his confidence very far at all:

> Here, many people feel that I should not reduce my confidence in $43 very far at all. …This intuition — which to a large degree I share — seems to cut directly against Conciliationism, and particularly against Independence. (Christensen, 2011, 8)

But that conclusion can't come just from looking at the initial details of the Careful Calculation case. Nor can the conclusion come from just the further stipulated details of the case described after the initial descriptions. Nor can that conclusion come from just the analysis of the higher-order evidence of the protagonist and the antagonist. There is a symmetry here between the epistemic situation of the protagonist and that of the antagonist. Both use the same following privileged self-knowledge claim as higher-order evidence confirming their own first-order evidence: "I know that I am paying very careful attention." The way Christensen sets up his response, it appears that the antagonist would also say that she knows that she doesn't have mental problems, isn't on drugs, and is sincere, and that she doesn't know the same about the protagonist's situation. There has to be an added inference from the fact that each has the same privileged self-knowledge claim to the conclusion: "I should not reduce my confidence" (Christensen, 2001, 8, quoted in context just above). There must be some symmetry breaker, some way of favoring one's own privileged self-knowledge claim over the other.

Christensen deploys the following symmetry breaker to favor his own "I know that I am paying very careful attention" over the other's:

> **Basing reason of inscrutability for favoring one's own privileged attention claim**
> "I can be very sure that I was actually paying attention rather than going through the motions of checking my answer, I cannot be nearly so sure that may friend was." (Christensen, 2011, 9)

This is Christensen's rational basis for favoring his privileged attention claim over the antagonist's same privileged attention claim. He can weigh more heavily his own claim to know he was paying attention over the antagonist's same claim because he knows his own attention status whereas he really can't know his opponent's. Christensen is here assuming that one's attention status is only something one can know. He is assuming that one can't know the attention status of the other disputant because, as privileged, it is inscrutable and so inadmissible as evidence. Someone else's attention level is inscrutable for everyone except the one who asserts a privileged claim to high attention.

Christensen uses the same basing reason for other cognitive features of Careful Calculation in order to affirm that one doesn't need to reduce one's credence level in this case. As we saw, he talks about not being able to count his opponent's equivalent claim not to be on drugs, not to being insincere, etc. because they aren't able to be scrutinized; they are unknowable, or at least significantly less known, for those who don't have this privileged self-knowledge claim, since they are privileged evidence. They are, thus, inadmissible as evidence against one's own privileged claims. I have formulated a condensed version of Christensen's assumption in the Inscrutability Thesis below, and you can compare it against the passage below where he best asserts it (Christensen, 2011, 11):

Inscrutability Thesis (a paraphrase of what is said in Christensen, 2011, 11)

> One can't know the privileged self-knowledge claims of others, and this fact is a good reason to take one's own evidence more seriously than the opponent's in a peer disagreement where every day, non-exotic mistakes have been eliminated by multiple reliable methods, and where the only possible mistakes at play are exotic ones which each is able to rule out only for themselves, since one can know that one has ruled out the exotic mistakes at play and one can't know the same about the other disputant's exotic mistakes.

What he says in (Christensen, 2011, 11)

> Consider how one would explain my friend's expressed disagreement in Mental Math. I know that doing a problem once in my head is not an extremely reliable process, because people commonly make undetected slips in mental calculation. So the overwhelmingly likely explanation for our disagreement obviously lies in one of us making this everyday sort of slip. Unfortunately, my personal information does not help me to eliminate this possibility for myself. Of course, there are also the exotic possibilities considered above: that one of us is tripping, psychotic, joking, lying, etc. And my personal information does allow me to eliminate various exotic possibilities for myself and not for my friend. But since these exotic scenarios are so unlikely, the fact that I can eliminate some of them has only a tiny effect on the plausibility of explaining the disagreement in a way that involves the falsity of my friend's claim. That is why I should (in categorical terms) suspend belief, or (in graded terms) come close to splitting the difference with my friend, in the sense of seeing the two answers as about equally likely to be correct. In Careful Checking, by contrast, the high degree of rational confidence I have in my initial belief is correlated with my rationally taking my reasoning method to be extremely reliable. And it is the extreme reliability of this method, a method which eliminates the "everyday mental slip" explanation of our disagreement, which both makes this sort of disagreement so unusual and makes the exotic explanations vastly more probable, should a disagreement occur. (This is why it's only in these cases that I'll think that something screwy must be going on.) At this point, when personal information allows me to eliminate several exotic possibilities for myself, but not for my friend, the balance of probability is shifted dramatically over to explanations involving the falsity of my friend's expressed belief. Thus it turns out that my high degree of initial rational confidence is correlated with my legitimately maintaining my belief in certain cases. It's correlated because when high initial confidence is appropriate, one generally may take one's reasoning method to be extremely reliable, which in turn eliminates everyday explanations for the disagreement, and makes exotic explanations—which tend to be sensitive to personal information—much more probable.

We will see how Christensen thinks that he has satisfied this Inscrutability Thesis. But we will show how he hasn't. The other scholars Christensen refers to who also prescribe steadfastness for Careful Calculating implicitly rely on this Inscrutability Thesis, and they also fail to satisfy the Inscrutability Thesis. There are two ways in which the Inscrutability Thesis fails, one because it is extremely hard to satisfy, and another because its fundamental assumption of inscrutability is false. In this criticism of Christensen here, we will initially

assume this Thesis is correct, and we will talk about how the higher-order evidence Christensen cites isn't adequate, given the fragile nature of the privileged self-knowledge claims to careful attention, as a symmetry breaker. Here Christensen, and others implicitly using the Inscrutability Thesis, falters because he doesn't see adequately the fragility of privileged claims that makes the Thesis extremely hard to satisfy.

In the next chapter we will show how the Inscrutability Thesis doesn't work at all because it segregates the observational-interpretive method of attaining self-knowledge from the privileged access method. There is no other scholar of peer disagreement who uses this very important Inscrutability Thesis to the extreme as Michael Bergmann. Christensen seems to be saying that one might be able to know a little bit about the other's privileged self-knowledge claim through observational methods in the following quote, though the person with the privileged self-knowledge claim has a tremendous advantage:

> "I can be very sure that I was actually paying attention rather than going through the motions of checking my answer, I cannot be nearly so sure that may friend was." (Christensen, 2011, 9)

But, Bergmann bases his deepest understanding of peer disagreement on the absolute segregation of the two methods. We will show the fragility of this Thesis due to the fragility of privileged self-knowledge claims. In Part One we saw how one can know things about the other and oneself through observation; Chapter Three and this chapter give more details about how one can know things about the other and oneself through observation. In the next chapter, Chapter Five, we show how extreme segregation of the two methods is misguided. Only in terms of the absolute segregation of the two methods engaged by Bergmann can we best present the remedy for such segregation, the Indirect Scrutability Norm. This remedy for the Inscrutability Thesis, gained through critique of its use in Christiansen's and especially Bergmann's approach to disagreement, is key to the foundation of the Assessment Framework developed in Chapter Six.

Getting back to the first criticism of the Inscrutability Thesis which argues the Inscrutability Thesis isn't satisfied, the first step in Christensen's argument for staying steadfast in the Careful Calculation case is to acknowledge that something weird is going on. It is weird that two people should disagree about such an easy math problem when they are equally capable in math skills, equally confident in paying attention, and equally use different reliable processes of calculating getting the same opposing numbers. Some exotic reason would

be needed to account for the discrepancy. Christensen immediately focuses on the following exotic abnormal things that have to do with attention and sincerity (Christensen, 2011, 8-9):

- Bizarre mental malfunction
- He is not sincere
- Actually exhausted
- Drunk
- Tripping
- Experiencing a confusing psychotic episode
- "Really only managing to go through the external motions of recalculating and checking, without actually paying clear attention"
- Joking
- "Messing with the other's head for fun."
- "Deliberately making false claims about his or her answer"
- A psychological or philosophical experiment
- "Problematize the hegemony of phallogocentric objectivity by an act of performance art."

The second step in Christensen's argument for being steadfast in the Careful Calculation case is to find that you can rule out all the exotic mistake possibilities for your own person. And Christensen does just this.

Yet, Christensen can't use the Inscrutability Thesis because it requires that there are only exotic mistake possibilities left to explain the discrepancies between the two disputants in Careful Calculation. And Christensen thinks that for both disputants the non-exotic mistake possibilities have all been ruled out by extremely reliable calculation methods used and reused on both sides. But, in Christensen's Careful Calculation case study there clearly are still some relevant and non-exotic simple mistake possibilities of attention that can't be detected and ruled out by the extremely reliable calculation and recalculation methods; we will see how this is the case below. Consequently, for the Inscrutability Thesis to work for returning a steadfast verdict, the one doing the reliability assessment must be able to detect any relevant mistake possibilities, whether they are exotic or not. And only each for herself can rule out her own exotic mistake possibilities. And if you can't rule out all the non-exotic relevant mistake possibilities on one's own side, then you can't favor your own position not knowing whether your opponent can rule them out as well.

We can see why it is important that there not be non-exotic mistakes possibilities for the Inscrutability Thesis to work in the following passage comparing the Careful Calculation case with only exotic mistake possibilities and the Mental Math case where there possibly are both exotic and non-exotic ones.

> Consider how one would explain my friend's expressed disagreement in Mental Math. I know that doing a problem once in my head is not an extremely reliable process, because people commonly make undetected slips in mental calculation. So the overwhelmingly likely explanation for our disagreement obviously lies in one of us making this everyday sort of slip.

> Unfortunately, my personal information does not help me to eliminate this possibility for myself. Of course, there are also the exotic possibilities considered above: that one of us is tripping, psychotic, joking, lying, etc. And my personal information does allow me to eliminate various exotic possibilities for myself and not for my friend. But since these exotic scenarios are so unlikely, the fact that I can eliminate some of them has only a tiny effect on the plausibility of explaining the disagreement in a way that involves the falsity of my friend's claim. That is why I should (in categorical terms) suspend belief, or (in graded terms) come close to splitting the difference with my friend, in the sense of seeing the two answers as about equally likely to be correct.
>
> In Careful Checking, by contrast, the high degree of rational confidence I have in my initial belief is correlated with my rationally taking my reasoning method to be extremely reliable. And it is the extreme reliability of this method, a method which eliminates the "everyday mental slip" explanation of our disagreement, which both makes this sort of disagreement so unusual and makes the exotic explanations vastly more probable, should a disagreement occur. (This is why it's only in these cases that I'll think that something screwy must be going on.) At this point, when personal information allows me to eliminate several exotic possibilities for myself, but not for my friend, the balance of probability is shifted dramatically over to explanations involving the falsity of my friend's expressed belief.
>
> Thus, it turns out that my high degree of initial rational confidence is correlated with my legitimately maintaining my belief in certain cases. It's correlated because when high initial confidence is appropriate, one generally may take one's reasoning method to be extremely reliable, which in turn eliminates everyday explanations for the disagreement, and makes exotic explanations — which tend to be sensitive to personal information — much more probable. But none of this undermines Independence. For in adjudicating explanations for our disagreement in any of these cases, I do not rely on my reasoning about the disputed matter. (Christensen, 2011, 11)

If there are non-exotic mistakes possible in the Mental Math case study, then one would not be able to rule them out on one's own side. And if one isn't able to rule them out on one's own side, then one can't be confident in higher-order evidence that one would know whether one is mistaken. And if one can't be confident in this higher-order evidence, then one can't leverage it to inspire steadfastness in the specific case thinking at least one knows one has higher-order evidence when one can't have comparable higher-order evidence that the other can rule out relevant mistake possibilities because the other person's primary evidence is inscrutable.

Christensen is simply wrong about there not being relevant non-exotic mistake possibilities of attention in Careful Calculation. He calls these in the quote just above "everyday mental slips." But think about a restaurant situation with all the noise and distractions. Suppose you read the printed total on the check and before the total gets fixed in your mind a friend distracts you by calling your name such that you make a mistake about one of the numbers; or suppose you did have what Christensen calls an everyday mental slip. You see the total on the check, and you accidentally remember it slightly the wrong way getting one of the digits wrong. Is that something that you would be able to detect with even extreme confidence that you were highly attentive? Maybe, but maybe not. Every one of these mistakes would account for the fact that he redoes several times all the calculations even with a calculator getting the same results. He simply in the quote above is wrong to think that the carefulness, repetitiveness, and multiple reliable methods used can rule out "everyday mental slips," even for the best of us, and especially in a loud and distracting context of a restaurant. None of those mistake

possibilities are exotic. They all seem relevant, not overly unlikely. Christensen is right, there is something unusual going on. But it is unwarranted for him to think his high confidence in high attention would detect all the subtle, everyday mistakes when you know something abnormal is happening and a cognitive peer also thinks she has a privileged high confidence claim and a privileged high attention claim.

> "My friend assures me that she did her calculations slowly and carefully, felt clear while doing them, and got her same answer repeatedly." (Christensen, 2011, 8)

After all, we have established that everyone deserves their privileged claims to be taken prima facie true unless there is reason to contest them. His following statement is unwarranted:

> "I can be very sure that I was actually paying attention rather than going through the motions of checking my answer, I cannot be nearly so sure that may friend was." (Christensen, 2011, 7)

Christensen's Inscrutability Thesis fails in the Careful Calculation case study to return a steadfast verdict as it proposes, and it is false by his own standards specified for the Thesis.

Christensen has made a mistake in applying the Inscrutability Thesis to the Careful Calculation case study just because he is not aware in this set up of the relevant and non-exotic fragility of his privileged self-knowledge claims to high confidence and high attention. Because of this fragile nature it is extremely hard to satisfy the Inscrutability Thesis, and Christensen clearly doesn't. In the next chapter we will see that the Inscrutability Thesis itself doesn't reflect the reality that the two methods are integrated with mindfulness. We will see that Christensen and others have made a deeper mistake in thinking that privileged self-knowledge claims are inscrutable, and we will remedy this mistake with the Indirect Scrutability Norm.

"I know that I have a headache," Sosa and initial substance and hidden, effective evidence

Ernest Sosa has two very interesting ways of motivating the steadfast view. Steadfastists generally focus on pointing out how there can be decisive and exclusive evidence that can justify staying steadfast even when one is disagreeing with a cognitive peer. He develops his two views for steadfastness in direct response to Christensen's Independence Thesis. Sosa thinks we don't need the Independence Thesis for two main reasons both having to do with exclusive evidence. First, though reasons or other sources of support for a belief often are hidden, they still can effectively support one's belief making the independent higher-order evidence Christensen needs irrelevant. Second, the initial substance of a disagreement can provide a basis for

downgrading one's opponent's epistemic status such that no independent higher-order evidence is needed (Sosa, 2010).

Sosa motivates one of his steadfast views with his understanding of hidden but effective evidence. The Independence Thesis presupposes that we need definitive higher-order facts about reasoning, FARs, in order to gauge whose first-order reasoning is the most reliable in a peer disagreement such that one can estimate whose view is more likely true. And we have to know what the FARs are in order to complete the comparative reliability assessment. Neither is so for Sosa. The epistemic state of a human being is such that often, especially in deep disagreements, key evidence, reasons, cognitive skills, or other sources of support are hidden or undisclosable, and even so they can effectively support one's beliefs. These hidden sources of support give one an edge in the disagreement.

Sosa's paradigm example of this effective, though hidden, evidence involves the knowledge one has that there are stars in the sky. He points to the fact that there are many different processes that hold that knowledge firm in one's possession, many of which were initially clear to one though many now are hidden:

> Why do we think there are stars in the sky? Can we really cite the reasons that led us to form and retain this belief? Sure, it may be thought, we know we've seen the stars repeatedly; that's our reason. But is that our only reason? And, more importantly, what about this proposition itself, that we have so seen the stars? What is our reason now for believing it? That we ostensibly remember it? We cannot require here the experiential memory of having seen the stars on a certain occasion, for this may well be missing. What is said to be enough is rather the ostensible retentive remembering of the fact in question. But such a paltry reason is unlikely to exhaust our epistemic justification (even if it does sustain some minimal subjective justification) for believing that there are stars in the sky or that we really have seen them. Even supposing that there is a distinctive 'appearance of retentive memory', which is already dubitable, what is further dubitable is that such an appearance could do much epistemically for our belief in the stars in the absence of the relevant past experiences that prompted this belief, a belief then kept in place through the proper operation of retentive memory. It is this latter, more substantial, time- and memory-involving rational basis that need not now be present to our reflective gaze in order to do its proper epistemic work. And so it is, I submit, for nearly the whole of one's body of beliefs. The idea that we can always or even often spot our operative 'evidence' for examination is a myth.
>
> If we can't even spot our operative evidence—so much of which lies in the past and is no longer operative except indirectly through retained beliefs—then we cannot disclose it, so as to share it. (Sosa, 2010, 291)

The point of this example is to show that there are crucial epistemic processes working in the background that we don't have access to needed for the justification of the belief that there are stars in the sky. Let's paraphrase Sosa's reasoning here that supports the idea that evidence hidden from one is also effective in justifying one's belief. Did your parents tell you initially about what a "sky" or "star" are, your teacher, your siblings? That is hard to say for anyone. You will think that you formed that belief because you have observed the stars in the sky repeatedly. Maybe that is good for a start. Yet we need more warrant than this. We need support for

accepting the proposition "There are stars in the sky." Apart from past experience of stars and the sky, what is the epistemic support right now for believing stars are in the sky? The memory of a specific time when you saw the stars in the sky. I can think of times that I have observed seeing stars in the sky. But there probably are people who think they did observe the stars in the sky but can't remember the specific occasion. So, it seems we need a vague appearance of a memory retaining the earlier observation of stars in the sky. In other words, a belief that there are stars in the sky is kept in place by a vague appearance of a retentive memory. But that vague appearance isn't strong enough to support the thought now that the stars are in the sky. We also need the assurance that there are stars in the sky based on past experiences. Yet that experiential temporal reasoning based on actual experience no longer is present for us today. And that experiential assurance is working in the background hidden to us helping one retain the belief that there are stars in the sky.

We can understand Sosa's hidden-but-effective view by understanding that he is an externalist about justification and a virtue epistemologist. These two orientations mean two things. One, nothing one is aware of is needed for justification. Two, justification of a belief is generated by, and held in place by, a virtuous disposition which is virtuous just because it embodies intellectual virtues aimed at true beliefs. It is the work of the intellectual virtues that ultimately justifies beliefs. They promote dispositions that are aimed at the truth.

These two epistemic orientations can be seen as operative in his way of handling peer disagreements with hidden-but-effective evidence. From Sosa's externalist virtue epistemology, reliance on hidden-but-effective evidence in peer disagreements makes sense. One's disposition embodies intellectual virtues aimed at truth. One's intellectual virtues are always operative also behind the conscious scene aiming at true beliefs rather than false ones. They are what justify beliefs, not any reason one can cite. So, it makes sense that one has effective evidence hidden from one's conscious awareness and generated from intellectual virtues embodied in one's disposition. Evidence can be hidden at the ready effective for getting at the truth of the matter in a peer disagreement even when one doesn't know what the evidence is.

The second argument for a steadfast stance in response to Christensen's Independence Thesis says the initial substance of a disagreement can provide a basis for downgrading one's opponent's epistemic status such that no independent higher-order evidence is needed (Sosa, 2010). This also makes sense given Sosa's

externalist virtue epistemology orientation. Intellectual virtues are operating behind the conscious scene embodying those virtues in one's disposition at the ready to generate in our everyday and intellectual lives beliefs that are more likely true than false. So, it makes sense that the initial thinking in a peer disagreement is all that is needed in a peer disagreement for staying firm in one's belief even if the other is an epistemic peer. After all, if the intellectual virtues upon which reliabilist justification is based were initially embodied in one such that they predispose one to true beliefs, then it is appropriate to go with the substance of one's initial thinking in the peer disagreement.

Sosa illustrates this initial-substance argument with the following case study:

> **"I have a headache"**
> Suppose you have a headache. What reason have you for thinking that you do? The important reason is, quite plausibly, simply that you do! Is this a reason that enables you reasonably to sustain your side of a disagreement when an employer believes you to be a malingering faker, with no headache at all. If so, then you can after all demote an opponent by relying on the substance of your disagreement. A huge part of your reason for rejecting the employer's claim that you're faking it is the very fact that gives content to your belief, the fact of the headache itself. Here then one has a conclusive reason that makes one's belief a certainty, even if that reason will be useless in a public dispute. It will not much advance your cause to just assert against your employer that you do have a headache, even if this is in fact the reason that makes you certain that you do. (This gulf between private and public domains will come to the fore again in due course.). Nor are headaches special in this regard. The same point applies to any obstreperous enough mental state. Anyone who denies to you that you are in that mental state is in a position like that of the employer who accuses you of faking it. Other examples of the same epistemic phenomenon will be found in any case of the given, whether it is the phenomenal given, such as our headache, or the rational given, such as the simplest truths of arithmetic, geometry, or logic. Here again if someone denies what you affirm, you can uphold your side by appealing to the very fact affirmed. Thus, if someone claims 2 and 2 not to equal 4, or a triangle not to have three sides, we could reasonably insist on what we know. (Sosa, 2010, 286-7)

The first thing to notice here is that this case in no way is an example of a peer disagreement. Each sees the other as having inferior epistemic status. The supervisor thinks the claim to a headache is insincere, and so thinks the person claiming it is wrong. And the person with the privileged self-knowledge claim, "I have a headache," thinks the employer simply is in no position to claim he doesn't have a headache. Of course, the employer thinks he is faking it, but only he has "a conclusive reason that makes one's belief a certainty, the fact of the headache itself." Sosa claims here that the initial substantial evidence is adequate for staying confident in one's belief. Here is the central privileged self-knowledge claim in Sosa's illustration:

> **Used to declare privileged self-knowledge of a particular mental state (Sosa):**
> 2) "I have a headache."

The way that Sosa understands the headache case study points to a way that privileged self-knowledge claims are fragile even from within a virtue reliabilist perspective. The disposition that Sosa expresses towards the "I have a headache" doesn't adequately express the fragile nature of this privileged self-knowledge claim.

Of course, it is very unlikely that someone claiming sincerely to have a headache actually doesn't have a headache. But this sort of thing does happen frequently enough that there is a category of mental health problems covering it in the DSM-5 handbook used by mental health providers for diagnosing mental health problems, which we will find described more below.

Sosa says the fact of the headache gives adequate justification for rejecting the view of the claim-critic; he says that this fact makes the belief the claimant has a headache a certainty.

> A huge part of your reason for rejecting the employer's claim that you're faking it is the very fact that gives content to your belief, the fact of the headache itself. Here then one has a conclusive reason that makes one's belief a certainty, even if that reason will be useless in a public dispute. It will not much advance your cause to just assert against your employer that you do have a headache, even if this is in fact the reason that makes you certain that you do. Nor are headaches special in this regard. The same point applies to any obstreperous enough mental state. Anyone who denies to you that you are in that mental state is in a position like that of the employer who accuses you of faking it. (From the quote just above)

Of course, Sosa is right, if it really is a fact that the employee has a headache, then the belief the claimant has of having headache is absolutely certain. But the question is, rather, whether it is indeed a fact that the claimant has a headache. Sosa's view here doesn't reflect the extremely low level of fragility of this claim. The reality is that sometimes, admittingly highly infrequent, people both think that it is a fact that they have a headache and think that the fact gives content to their belief that they have a headache; but they really don't have a headache. Psychologists recognize that people sometimes fully think they have a headache when they really don't; an example is people with Specific Phobia DSM-5 300.29 and an even more specific phobia described in the International Classification of Headache Disorders, that is ICHD-3, in its designation A12.6.

Sosa's way of talking about "I have a headache" case doesn't reflect the actual reliability of that privileged self-knowledge claim. It is highly reliable that a person who sincerely says she has a headache really does have a headache; but not certain. Sosa doesn't acknowledge this fragility of the claim "I have a headache." He further says that nobody who says you don't have a headache when you sincerely say, "I have a headache," can be a peer: "Anyone who denies to you that you are in that mental state is in a position like that of the employer who accuses you of faking it." The position he is talking about here is that of a person who has been demoted from peerhood on this matter (Sosa, 2010, 286-7). He might say that it is so improbable that one has DSM-5 300.29 such that there is no need to take into account such an improbable possibility. That would be the case normally. But, if one is having a disagreement with someone who is a strong peer, that peer would fully

understand the improbability of having such a DSM-5/ ICHD-3 diagnosis, and that person would have good considerations for thinking the claimant really doesn't have the headache.

To see this, consider a similar case study but this time with a peer denying the person has a headache:

> **HEADACHE**
> Kai calls his boss telling her he can't work today because he has a severe headache. The boss says, "No you don't." Kai is incensed and says, "How could you accuse me of faking that I have a headache? I know that I have a headache!" Kai's boss says, "No I am not accusing you of faking it. You really do sincerely believe that you have a headache. Remember when you told me you have an official diagnosis of DSM-5 psychological problem called Specific Phobia DSM-5 300.29? I did some research on this even consulting my own psychologist after you called in last time with a headache, and I found out that headaches are a common symptom of DSM-5 300.29. In fact, you seem to have all the criteria of what the International Classification of Headache Disorders, that is ICHD-3, calls a headache attributed to a specific phobia A12.6. Those criteria are DSM-5 300.29 with a headache that occurs exclusively when the person is anticipating exposure to a phobic stimulus, and you have told me several times that working on the top of the building site gives you fear and nausea. If that makes sense, you could come in when you are feeling better, and we could figure out how to put your work site on the ground." Kai sees this higher-order evidence as a huge threat to his privileged claim. He realizes that it is rare for anyone to contest someone's sincere privileged claim to having a headache. He still thinks he has a headache. But, now he thinks each is about as likely to get the issue right. He wants to do some of his own research on this before he is willing to think about giving up on his privileged claim.

The employer is a strong epistemic peer in this disagreement. Of course, this kind of situation is rare. But sometimes a sincere statement "I have a headache" doesn't reflect the reality of one's mental state. The claim "I have a headache" doesn't certainly reflect the reality of the person's mental state.

Notice in HEADACHE the strong peer fully recognizes that it is rare that someone claiming sincerely to have a headache really doesn't. The boss continues the disagreement because she finds observational evidence that the situation isn't normal. The boss observes Kai's behavior and listens carefully to the way Kai describes the circumstances of his headache. She has observation evidence that isn't consistent with having a headache. And she finds observational evidence of a phobia. She is able to challenge Kai's privileged self-knowledge claim by pointing out observational information that runs counter to behavior of a person who really has a headache. Sosa doesn't acknowledge the fact that a peer can indirectly scrutinize a privileged self-knowledge claim through observation. And this kind of challenge of a strong peer is based on observation. This is an example of what we talked about in the last section with regard to the Inscrutability Thesis. Had he portrayed a strong peer in his headache case study, like the one just above, he would have run into this challenge and have to respond to it. Sosa seems to think that privileged self-knowledge claims are not scrutable for the people that don't have them. We see in Sosa, as we saw in Christensen, the lack of awareness that a peer can legitimately

challenge privileged claims in the way we described, and in this he also doesn't see the fragile nature of such claims.

But the oversight here about the fragility of privileged self-knowledge claims points to the wider need from a virtue reliabilist perspective, or any other perspective, to be more aware in discussions of peer disagreement of the fragile nature of privileged self-knowledge claims. For the virtue reliabilist the justification of a belief is generated by, and held in place by, a virtuous disposition which is virtuous just because it embodies intellectual virtues aimed at true beliefs. Privileged self-knowledge claims are fragile because of the many psychological, social, and physical barriers to their success. We ought, therefore, to foster epistemic dispositions that reflect the fragility of them. Any evidence that is hidden is evidence in an unconscious context. And, as discussed in Part One, even the best information or evidence that one has stored and hidden is subjected to the many unconscious ulterior motives not aimed at the truth about one's mental states; but, rather, aimed at only making one think better about oneself.

"I have decisive personal information" and Lackey's personal information

Jennifer Lackey focuses on a different kind of disclosable yet privileged evidence. She talks about personal information as something that she has and nobody else has (Lackey, 2010, 310). Personal information is information about oneself only the individual person can know about, for example, how one's cognitive capacities are, whether one is not suffering from depression, having side effects from medication, exhaustion, or distraction. One can share these things with others, but only the individual knows them with first person authority.

Personal information plays a decisive role in her view she calls "justificationist," which says, "the amount of doxastic revision required tracks the degree to which the target belief is confidently held and highly justified" (Lackey, 2010, 277). Her view on peer disagreement, thus, is laser focused on the level of justification disputants have, and one must revise their attitude towards a belief according to the level of justification. And personal information is at the heart of the justificationist perspective. In fact, it says, the only way to retain the same degree of belief is if one has such personal information "that provides a relevant symmetry breaker"

(Lackey, 2010, 277). Personal information is the higher-order evidence that breaks the symmetry allowing one to retain the same credence. Thus, for the justificationist view to work and one is able to retain one's credence, one must be able to assert, either explicitly or implicitly, the following privileged claim that counts as high-order evidence that breaks the symmetry of justification between two peers:

Used as higher-order evidence confirming first-order evidence (Lackey):

3) "I know I have decisive personal information."

Without this privileged self-knowledge claim, without the dimorphism of the personal information it affords, each should seem about as likely to get it right.

One of Lackey's case studies and her assessment of credence level resulting from it show that she is not adequately aware of the fragile nature of privileged self-knowledge claims. Here is that case study:

> *Bird*. While reading in the library with my friend Eva, I glance out the window, catch a glimpse of a bird flying by, and on this basis hastily form the belief that a magpie just flew by. After saying to Eva, who was looking out the window at the same time, that I enjoyed seeing the magpie that just flew by, she responded, "Nothing flew by the window." Prior to this disagreement, neither Eva nor I had any reason to think that the other is evidentially or cognitively deficient in any way, and we both sincerely avowed our respecting conflicting beliefs. (Lackey, 2010, 287)

Lackey concludes that the protagonist has no pressure here to reduce confidence because she has decisive personal information that gives here the upper hand in this peer disagreement reasoning as follows:

> If I look out a clear, nearby window, have my contact lenses in, have not been drinking or taking any drugs, and know all of this to be true of myself, then Eva claiming that nothing flew by the window does not seem to give me any reason at all to revise my belief that it was a magpie that flew by. Instead, given that Eva was looking out the window at the same time as me, such a disagreement seems appropriately regarded by me as evidence that something is not right with her, either evidentially or cognitively. … What is it about the nature of the disagreement in this case that renders it intuitively rational for there to be no doxastic revision on my part? At least prima facie, one answer is that the disagreement in question seems outrageous. Sure, the bird may have instead been a starling, a grackle, or a red-winged blackbird, and so I may be wrong in my belief that it was a magpie. It may also be possible, though very unlikely, that it wasn't a bird at all that flew by but, rather, was an extremely large insect or a bat that forgot it was nocturnal. Disagreements of either of these sorts—i.e. regarding whether it was a magpie or a bird—elicit quite different intuitions regarding doxastic revision. But as Bird is described, *something* clearly flew by and so Eva's disagreement with me on the grounds that nothing flew by seems so outrageous that it lacks the epistemic significance it might otherwise have had.
>
> It is not difficult to see, though, that the explanation at work here does not bottom out in mere outrageousness. For what explains the outrageousness of the disagreement in question is that I have an extremely high degree of justification in my confident belief that something flew by the window. … The very high degree of justification had by this belief [namely, that something flew by the window] enables me to downgrade my opponent's epistemic status in the same manner found in Directions. … Once again, my view can handle this problem [that is, a claim of Eva's in a different scenario.] with ease … In particular, my belief that a medium-sized black and white flying creature flew past the window is protecting belief, one that is also challenged by the disagreement in question but is one for which I have a very high degree of justification. (Lackey, 2010, 287-9)

Because she doesn't adequately recognize the fragile nature of her privileged claim to personal information, Lackey is simply wrong that in this case study there is no pressure for the protagonist to reduce confidence. Very similar to what we saw Christensen do in his analysis of Careful Calculation, and actually similar to what many scholars of peer disagreement and people in everyday life generally do, Lackey overlooks non-exotic mistake possibilities due to the fragile nature of the prima facie true status of privileged claims to decisive personal information. These everyday common mistakes that can occur are not obvious, rather subtle and not easy to see, as we saw when talking about Christensen's work. But in not looking out for these common but not obvious mistake possibilities, one improperly overestimates peer credence levels, or under-estimates, in a peer disagreement, and one epistemicly inappropriately estimates the likelihood that one will get the issue right.

So, let's talk about relevant mistake possibilities she would need to be able to rule out if she is going to claim what she actually claims in this case study. There are two that I can think of. One, what looked like something flying could have been what is called a floater by ophthalmologists. They are the result of aging. They are black and white, just as she described the thing in the window. They appear to fly off quickly, just as she described the thing in the window. If she is going to be so dismissive of her friend Eva's claim that there was nothing flying in the window, so confident that she is right, surely she would need to rule out this everyday alternative explanation of what happened. Instead, she tends to think her friend has some exotic problem and demotes her friend's epistemic status. Apart from the very real practical possibility of losing a friend who rightly thinks the protagonist inappropriately doesn't recognize such everyday mistake possibilities, apart from the fact that her oversight creates an inappropriate assessment of the other's epistemic status, the protagonist loses a possible opportunity to gain more personal information about her eyesight. Sensitivity to the fragile nature of privileged self-knowledge claims is an epistemic virtue.

Certainly, the protagonist should rule out non-exotic and common mistake possibilities before she resorts to the exotic ones like thinking that her friend is cognitively impaired. Here is another relevant and non-exotic mistake possibility the protagonist would need to rule out: A natural optical illusion projected on the window. I have experienced many times such natural illusions. They can happen when the light behind one reflects the image of something also in back of one onto a window in front of one. I have also seen light interacting with a moving object on the street, like a car or a person, casting that image on a window in front of me. Once an image of my son seeming just outside the window looked so realistic that I ran to the window only to find my son was not there, but rather was coming in a door at the other side of the coffeehouse.

Perhaps Lackey isn't familiar with floaters or such natural optical illusions. But, that is just the point. People are often not adequately aware of the fragile nature of privileged self-knowledge claims to personal information. She says in this situation that she has "extremely high degree of justification in my confident belief that something flew by the window." But, she really doesn't have an "extremely high degree of justification" if her justification can't rule out these two non-exotic everyday mistake possibilities, or any other such relevant non-exotic mistake possibilities. So here, not only does the perspective the case study was designed to illustrate promote a mistaken assessment of epistemic status of the antagonist, but also it promotes a mistaken assessment of the epistemic status of the protagonist.

It also appears that Lackey's assessments of both disputants harmonize with the Inscrutability Thesis we have found explicit in Christensen's view of peer disagreement and implicit in Sosa's. Lackey would itemize many of the items of personal information she has that are relevant to this case study, she hasn't been drinking, she hasn't taken drugs that can make one hallucinate, she knows her contacts are in, etc. Yet that list in this situation should include knowledge of the status of her eyes as regards floaters and knowledge of one's surrounding as regards anything that can project a natural illusion. Were Eva really an epistemic peer she would also know that it is unlikely that her friend would think there is a bird flying outside the window when there really isn't. And such an epistemic peer would likely bring up observational information she has about the

protagonist that counts as evidence that her friend and not her, i.e., Eva, is the one that made the mistake. Eva could point out that she has seen her friend looking sideways for something that apparently isn't there, which is what people with floaters do when they start. Perhaps Eva could point out her observations of the surroundings that her friend finds things that are conducive to natural illusions. Dismissing the view automatically and demoting them without listening to their concerns and observational evidence isn't something one normally does when they are having a disagreement with a truly epistemic peer. The point here is that Lackey, through this case study, inappropriately doesn't seem to acknowledge that there can be observational evidence that could indirectly scrutinize a person's privileged self-knowledge in the form of personal information. Instead, the view of peer disagreement she recommends doesn't adequately recognize that personal information can be indirectly scrutable which means the Inscrutability Thesis is false. Of course, Lackey knows that one can be wrong about one's perceptions. But, her view seems to be that once one has come to feel sure about one's personal information, that information can't be scrutinized. A strong peer is naturally going to push back against alternative evidence such as privileged self-knowledge in the form of personal information. Otherwise, the person really isn't a peer.

Bergmann, extreme confidence, and seemings

I will talk thoroughly about Bergmann's use of privileged self-knowledge claims in a way that defines the deepest parts of Bergmann's philosophical response to the skeptic about objects and perceptions and also peer disagreement in the next chapter; here I am giving just a summary of findings. I treat this in its own chapter because it is the only one of the five scholars who defines his overall philosophy in terms of his use of privileged self-knowledge claims, because it is very complex and needs more room, and because it so well represents the most challenging uses of privileged self-knowledge claims. The following is the privileged self-knowledge claim that Bergmann has to help people stay steadfast in a peer disagreement over moral issues or over any issue.

Used as primary evidence for a belief (Michael Bergmann)

4) "I am extremely confident that p based on 'I am extremely confident in my insight about the way things really are which implies p.'"

I give you here the results of the analysis of Bergmann's obliviousness to the fragility of the prima facie true status of privileged self-knowledge claims. He inappropriately uses privileged claims to extreme confidence to support beliefs he finds important, and this is so because he doesn't adequately take into account the fragile nature of "extreme confidence" claims. To find the argument for this view you will have to go to the next chapter. We will use these conclusions, though, in the summary conclusions of this chapter. Bergmann's views of peer disagreement express the most clear and extreme case of segregating the two methods of knowing oneself, the observational-interpretive methods. As we will see in the next chapter, Bergmann finds that privileged self-knowledge claims are segregated from observational-interpretive indirect scrutiny. And we turn to his work in the next chapter because we can give the best remedy to such segregation, the Indirect Scrutability Norm, only when we best see the problem the remedy is for.

"I know my base reasons for my belief, and I know they entail the conclusion." Kelly base reasons

We can learn much about Thomas Kelly's view of peer disagreement and his crucial privileged self-knowledge claim by describing how he is critical of Christensen's Independence Thesis. Kelly argues against the Independence Thesis by saying that it avoids what he calls the "burden of judgment" (Kelly, 2013). What he means by this is that making a judgment initially about one's epistemic status relative to the other disputant's in a particular disagreement is extremely difficult, a kind of burden in fact. Kelly's thought is that one's initial comparative judgment about one's standing in a peer disagreement can be sufficient when that judgment has been carefully deliberated with evidence and without begging the question such that there is no need for the post-initial-judgment facts of reasoning, FARs, which Christensen talks about. We will be discussing his thinking after 2013 where he advocates the Total Evidence view. When you take on all the heavy and onerous responsibilities in the process of making a fallible decision you think you have a good basis for making—realizing there will always be people looking on from their armchair making their own comparative reliability-assessment—this is what Kelly in "Disagreement and the Burdens of Judgment" calls the burdens of judgment:

> Faced with a peer who disagrees, knowing how one is rationally required to respond will typically require an extremely substantive judgment about one's overall epistemic situation, as opposed to the straightforward application of a general norm that dictates agnosticism in all such cases. Such are the burdens of judgment. (Kelly, 2013, 52)

One of the most important aspects of the burden of judgment is getting it right about what one's basing reasons are:

> For notice that, if in fact you reasoned impeccably in arriving at your original answer, then the facts from which you reasoned (that the total bill is n dollars; that m people have agreed to divide the check evenly, etc.) literally *entail* the correct answer. So if such facts are among the evidence you have to go on in evaluating my belief, then they would seem to provide a basis for discounting my opinion entirely. But according to Independence, you should set aside such facts when evaluating my belief. (Kelly, 2013, 38)

Notice there is huge stress on the basis upon which one makes one's conclusion. Here "the facts from which you reasoned" are the basing reasons, and those facts provide a basis for making a judgment, they "entail" the judgment. Christensen thinks the post-initial-judgement is needed in order to be fair to each and so avoid question-begging. Christensen thinks the Independence Thesis requires us to engage the comparative reliability-assessment of the initial judgement reasoning without weight given to the initial judgement, bracketing out the epistemic influence of the initial judgement. But, says Kelly, if the initial judgement is done well, it will avoid question-begging and bias.

We can now say that when engaging the central burden of judgment, Kelly focuses on the basing relation. Kelly is very intentional about focusing on the basing relation. There must be a good basing relation such that the facts in terms of which one argues entail the conclusion promoted. So, for Kelly, to engage the burden of judgement is to know one's base reasons for one's belief and know that the base facts or evidence entail the conclusion. The following privileged self-knowledge claim, thus, is needed for taking on the burden of judgment:

Makes rational deliberation possible (Kelly)

5) "I know my base reasons for my belief and I know they entail the conclusion."

Whereas the discussion of the four other scholars above pointed out mistakes that did lead to inappropriate assessments of the comparative epistemic status of disputants, I will here argue that Kelly's burden of judgment approach to the initial thinking in a peer disagreement doesn't adequately acknowledge two considerations that can lead both to inaccurate assessments of comparative epistemic status and to a theory of peer disagreement that doesn't account for the most challenging peer disagreements about privileged self-knowledge claims. First, Kelly doesn't give an example of strong peer disagreements when developing his position against Christensen's

Independence Thesis. Second, he doesn't see that a strong peer in a disagreement often uses the observational-interpretive method to appropriately challenge the claimant's initial privileged self-knowledge claim to know one's basing reasons. In using the observational-interpretive method to challenge the privileged self-knowledge claim to basing reasons, we verify the Indirect Scrutability Norm that will be crucial for deriving the Assessment Framework in Chapter Six.

> **Indirect Scrutability Norm:**
> If one disputant has significant or highly significant observational evidence that implies the other's privileged self-knowledge claims is questionable, then don't trust that claim according to the following credence levels: highest signs = lowest credence; significant signs = low credence; little signs = high credence; no signs = highest credence.

We will see in two of Kelly's case studies below how his burden of judgment approach to the initial thinking in a peer disagreement can lead to inaccurate assessments of comparative epistemic status.

A non-peer in a case study that illustrates the Burden of Judgment view

In the following case study about the Holocaust illustrating what he finds wrong with Christensen's Independence Thesis, Kelly presents a weak peer:

> **Holocaust Denier:**
> Suppose that I possess a great deal of evidence that bears on the question of whether the Holocaust occurred; I look it over and judge correctly that this body of evidence strongly confirms that the Holocaust occurred; on the basis of that assessment, I invest a correspondingly high amount of credence in the proposition. I then encounter a Holocaust denier. (For purposes of the example, let's imagine that this person is quite reliable when it comes to matters that are unrelated to the Holocaust.) In evaluating the epistemic credentials of his belief that the Holocaust never occurred, Independence would have me bracket my assessment of all of those considerations which led me to believe that the Holocaust did occur. An obvious question is whether, once I do this, I'll still have enough left to go on to offer an evaluation of the Holocaust denier's belief. A second question is why, even if I do have enough left to go on to arrive at an evaluation, we should think that the evaluation that I come up with under those conditions is worth anything.
>
> Suppose that the person in question is grossly ignorant of certain historical facts, historical facts which make it overwhelmingly likely that the Holocaust occurred. Indeed, perhaps the evidence that the Holocaust denier possesses is sufficiently impoverished and misleading (the misleading testimony provided by parents whom he had a default entitlement to trust; the propaganda to which he has been subjected, etc.) that his belief that the Holocaust never occurred is a perfectly reasonable thing for him to think, both objectively and by my own lights. His problem is not irrationality but ignorance. One might have thought that his gross ignorance is certainly something that I should take into account in evaluating the epistemic credentials of his belief. (Recall that, for reasons given above, evaluating a person's belief in the sense relevant to Independence must go beyond merely making a judgment about the epistemic status of his belief given his evidence.) However, there seems to be a problem with my doing this. Suppose that it turns out that (as is plausible enough) the historical facts of which he is ignorant are the very same facts on which I base my own belief that the Holocaust occurred. In that case, in evaluating his belief, I should bracket my own assessment of these considerations. That is, I should set aside my own judgment that these considerations strongly support the view that the Holocaust occurred. But the problem then is this: my judgment that the Holocaust denier is grossly ignorant when it comes to matters relating to the Holocaust is not at all independent of my assessment that the relevant considerations strongly confirm the occurrence of the Holocaust. That is, if I set aside my assessment that these facts strongly confirm the occurrence of the Holocaust, then I would no longer take someone's ignorance of them to be a handicap in judging whether the Holocaust occurred. After all, there are ever so many facts ignorance of which I take to be no handicap at all when it comes to judging whether the Holocaust occurred. It is only because I judge that these facts confirm that the Holocaust occurred, that I take ignorance of them to be at all relevant to "the epistemic credentials" of someone's belief about the Holocaust. (Kelly, 2013)

Kelly argues there is no way to appropriately solve this problem intuitively using Christensen's Independence Thesis. In initially solving the issue of the disagreement, the protagonist uses the historical evidence that he has. His basing reason is the substantial body of historical evidence he is aware of for the Holocaust. From the perspective of the Independence Thesis, the person cannot consider that evidence when there is a peer disagreement, since it was used in the initial evaluation of the judgement. So, there would be, according to Kelly, no way of saying that the Holocaust denier is more likely to be wrong. To which Kelly, of course, says that is wrongheaded, since it would not be appropriate to slight first-order evidence when that very well could be the best and sufficient evidence.

In the process of defending his criticism of Christensen's Independence Thesis, Kelly presents a weak epistemic peer. Right off the bat, Kelley stipulates that the antagonist is "quite reliable when it comes to matters that are unrelated to the Holocaust." Consequently, it appears that Kelly's Holocaust example really isn't a case of epistemic peer disagreement. Maybe he would be an epistemic peer about other matters, but certainly not about an issue related to the Holocaust. So, in this example, he isn't talking about peer disagreement. He easily demotes the antagonist. After considering one more of Kelly's case studies below, I will comment on both the cases studies Kelly uses to illustrate and support his burden of judgment approach.

A moderate peer in a case study that illustrates the Burden of Judgment view

Kelly illustrates and supports his burden of judgment view in another case study that involves two peers who are jurors having a disagreement over the innocence of the defendant:

> You and I are attentive members of a jury charged with determining whether the accused is guilty. The prosecution, following the defense, has just rested its case. Suppose further that neither of us has any particular reason to think that he or she enjoys some advantage over the other when it comes to assessing considerations of the relevant kind, or that he or she is more or less reliable about the relevant domain. Indeed, let's suppose that we possess significant evidence that suggests we are likely to be more or less equally reliable when it comes to questions of the relevant kind. Because we're aware of this, if we had been asked in advance of the trial which one of us is more likely to be wrong in the event of a disagreement, we would have agreed that we were equally likely to be wrong. Nevertheless, despite being (apparent) peers in these respects, you and I arrive at different views about the question on the basis of our common evidence. For example, perhaps I find myself quite confident that the accused is guilty while you find yourself equally confident that he is innocent.
>
> Suppose next that, upon learning that I think that the accused is guilty, you reduce your confidence in his innocence. However, even after you take my opinion into account, it still seems to you that on balance the evidence suggests that he is innocent. You still regard it as significantly more likely that he is innocent than that he is guilty, to the point that you can correctly be described as retaining your belief in his innocence. Question: in these circumstances, is there any possibility that this is a reasonable response on your part? … Suppose that the original evidence with which we are presented strongly supports the view that the suspect is innocent. Your original belief is a rational response to what was then our total evidence; mine is not. (Against a general background of competence, I commit a performance error.) After you learn that I think that the accused is guilty, your total evidence has changed: it is now on the whole less supportive of the view that he is innocent than it was previously. It is thus reasonable for you to reduce your confidence to at least some degree. Still, the total evidence

> available to you then might very well make it more likely that the suspect is innocent than that he is guilty, to the point that it's reasonable for you to believe that he is guilty. In any case, there is certainly no guarantee that the uniquely reasonable response on your part is to retreat to a state of agnosticism between your original opinion and my original opinion, as the Conciliationist suggests. (Kelly, 2013)

Kelly here describes a peer disagreement where the challenge to a person's privileged self-knowledge claim to basing reasons is moderate. Each has the same evidence. Each has equivalent skills for evaluating the evidence. They have an equivalent understanding of the background of the issue. In this case Kelly finds it reasonable for the juror who believes the defendant is innocent to reduce confidence when finding out a peer disagrees; but this juror can reasonably still stay steadfast to her conclusion the person is innocent when this juror's basing reasons really do entail the conclusion of innocence. Even in a peer disagreement like this, one can appropriately rely on the initial reasoning and the privileged self-knowledge claim about one's basing reasons; and this means Christensen's Independence Thesis is false. The claim-critic is a moderate peer because he doesn't continue to dig deeper by bringing up new considerations that challenge the claimant's privileged self-knowledge claim about basing reasons.

Three conditions for a strong peer

Before getting back to discussing Kelly's work, let's see where three general conditions for a strong peer fit into the existing views of peerhood in the literature. We have been talking about the first two below as general conditions for any peer disagreement:

The first two conditions together describe a peer, the third added marks a strong peer

1) Must have comparable understandings of relevant objects of knowledge relevant to the issue, like beliefs, details, evidence, or facts.
2) Must have comparable cognitive skills relevant for gaining and evaluating the objects of knowledge.
3) The higher level of comparables of 1) and 2) is such that one is assessed as roughly as likely to get it right.

The first two conditions together describe a peer. The third condition added to the first two altogether describe a strong peer.

In explaining the components of peerhood, scholars generally include, on one hand, the need for each disputant to possess highly comparable objects of knowledge (1)), that is, the things that one knows or claims to know, like evidence related to the issue, the weight of evidence, relevant beliefs, facts, information, details, etc.; and, on the other hand, the need for each to possess highly comparable cognitive skills (2)), that is, tools for attaining knowledge or for evaluating the objects of knowledge, like perception, logical reasoning skills,

working memory, etc. For example, David Christensen recognizes highly comparable skills of thinking and of evidence of those skills as signs of peerhood (Christensen, 2009, 758-9; 2018). Jennifer Lackey takes comparable information about the disputants' thinking as indications of peerhood (Lackey, 2010, 277-289). And Michael Bergmann describes peerhood as about evidence of highly comparable internal and external rationality (Bergmann, 2009).

While I acknowledge the value of attempts to evaluate the merits of each specific scholars'version of 1) and 2) in order to find the ones that reflect best the way peer disagreements work in real life, it is important to realize that different domains of peer disagreement often have cognitive skills and objects of knowledge exclusive to their domain. Scholars sometimes present their description of conditions for peerhood in a way that makes one wonder whether they think the conditions apply to all disagreements. For example, Bergmann thinks of evidence and cognitive skills in terms of what he calls "felt veridicality" and "insight" (Bergmann, 2009). And for Lackey one of the most important types of evidence is what she calls "personal information" (Lackey, 2010); such information can break the symmetry between disputants making one able to think one is more likely to get the issue right. I will be focused on the objects of knowledge and cognitive skills exclusive to peerhood in target disagreements.

All views about what it means to have a peer in a target disagreement, whether they be steadfast or conciliatory oriented, have or should have an understanding about when it is appropriate to see the other as roughly as likely to get the issue in question right. This is a version of Adam Elga's understanding of peerhood (Elga, 2007, 487), and David Enoch has similar views (Enoch, 2010, 956). For target disagreements, declaring there is never any need for 3) is unrealistic given the copious empirical studies discussed in Part One that demonstrate the psychological, social, and physical barriers to accurate privileged self-knowledge of one's mental states.

Alex Gelfert is right, this way of thinking about peerhood as "roughly likely to get it right" by itself leaves out the specificity needed to judge whether one is a peer (Gelfert, 2011). I am not guilty of that mistake because I use this way of talking about peerhood as merely a landmark demarcating the degree of 1) and 2) where a peer in a disagreement becomes a strong peer, that is, the point where the peer is roughly as likely to get it right; the

peerhood reaches this degree of peerhood when the comparability described in 1) and 2) are met to a higher degree. So, Gelfert is also right when he says, "epistemic peerhood comes in degrees" (Gelfert, 2011, 508). For all epistemologies of peer disagreement this landmark is reached when there is a critical depth of comparability.

Different scholars have different ways of expressing the conditions where each is roughly as likely to get it right, the critical landmark level of symmetry marking a strong peer. For Christensen it would seem if one doesn't have evidence that one's own reasoning is more reliable, one must reduce confidence in one's view (Christensen, 2009, 758-9; 2018). And this should make one come to think that the other is roughly as likely to get it right. The phrase "roughly as likely to get it right" is inspired not only by the work of Elga and Enoch, but also by Christensen (Christensen, 2007, 193-4). For Lackey it seems, I should lower my confidence when I don't have personal information about myself that implies my thinking on the matter is more reliable (Lackey, 2010, 277-289). As we will see, for Michael Bergmann if one is in a disagreement where there is equivalent internal and external rationality and one isn't confident that one's insight reflects the way things actually are, then one should think the other is roughly as likely to get the issue right (Bergmann, 2009).

Let me explain my use of "comparable" and "roughly as likely" in the description of 1) and 2) above, and also my use of a new phrase demarking the difference between a peer and a strong peer, "highly comparable." I take Nathan King (King, 2012, 263), Kirk Lougheed (Lougheed, 2020, 70), and Ernest Sosa (Sosa, 2010, 290) to be right, that there never, or rarely, is a time when people have exactly the same evidence and skills, or exactly the same likelihood of getting the issue right. With these three categories just above the demarcation between a peer and a strong peer can be seen. They allow one to see the other as a strong peer even when she doesn't share the exact same cognitive skills and objects of knowledge.

While an epistemic peer has comparable cognitive skills and knowledge objects like evidence or beliefs, a strong peer has these to a higher degree. The following is one way of distinguishing the higher degree in a disagreement where disputants are roughly as likely to get it right:

Signs of highly comparable
When my peer employs reasoning, which is foundational also for me, to successfully call into question my use of the cognitive skills and evidence that support my view, I assess that those skills and evidence now only make me slightly more likely to get it right than my peer; and I assess that my peer is now roughly as likely as me to get the issue right.

To understand this we can deploy our own distinction between first-order comparables and higher-order comparables. The first-order comparables consists of the initial skills and evidence I present to the other disputant. Higher-order comparables surface when the other disputant employes reasoning I too use as foundational to call into question my initial skills and evidence. Think of this as like two athlete runners disagreeing before a race about who will win. They acknowledge comparable training in relevant skills, knowledge, conditioning, and experience. One athlete thinks her conditioning is slightly better. She uses skills and evidence for thinking why her conditioning is slightly better. Her opponent uses reasoning that she deeply accepts to undermine and call into question the skills and evidence she used to support her claim that her conditioning is better. She now thinks she is roughly as likely to get the issue right, though she thinks she is slightly more likely to be right because she thinks some of her evidence still survives the undermining. While a peer doesn't have the higher degree of comparables that can challenge my view such that I think she is as likely to get it right, the strong peer does. I will illustrate how this works in the examples in Chapter Six.

Strong peers often challenge one's basing reasons using the observational-interpretive method

The problem with developing a theory of peer disagreements by illustrating it with a non-peer or a moderate peer—as Kelly is doing in the Holocaust and the juror case studies—is that the theory doesn't account for the unique challenges that surface with strong peers. The case study with the two jurors, while indeed a peer disagreement, isn't a strong one. The claim-critic doesn't dig deeper by presenting a concern that challenges the claimant's privileged self-knowledge claim about basing reasons. And if the challenges of a strong peer aren't adequately taken into account, inaccurate assessments of comparative epistemic status ensue. The unique quality of the strong peer is that she is in a position to appropriately challenge the claimant's initial privileged self-knowledge claim about basing reasons, since she is equivalently knowledgeable about the details of the issue, and equivalently capable of evaluating the issue. In both the non-peer and the moderate peer disagreements Kelly presents, there is no deeper challenge to the initially privileged self-knowledge claims about a basing reason. The strong claim-critic fully understands what it means to have a privileged self-

knowledge claim. He understands that people can have immediate, uninterpreted, and non-observational self-knowledge. Yet, he digs deeper and brings up new considerations that challenge the privileged self-knowledge claim about basing reasons, considerations the claimant had not adequately addressed. One way to challenge a privileged self-knowledge claim about basing reasons is to give evidence that the basing reasons don't adequately entail the conclusion the basing reasons are meant to support. A second way of challenging the privileged self-knowledge claim about basing reasons, which we are pursuing in this chapter, is to provide evidence that the claimant doesn't have the privileged self-knowledge claim about basing reasons that she thinks she has. If the claim-critic doesn't have considerations that bring the claimant's claim, that claim-critic wouldn't be an epistemic strong peer.

To be clear, we are not talking about a strong peer claim-critic asserting that the claimant's basing reasons don't entail the claimant's conclusion. Rather, we are talking about a strong peer claim-critic who asserts the claimant doesn't have the basing reasons she asserts she has.

We can see an example of strong peerhood if we extend Kelly's Holocaust case study with the Holocaust denier pushing back against the protagonist Holocaust affirmer who has the privileged self-knowledge claim that the base reasons of his view are the high-quality historical research studies he has read. Were the Holocaust denier a strong peer she would likely be able to point out challenging evidence that the Holocaust affirmer doesn't use the basing reasons he asserts he uses, but rather has different basing reasons. Say the claim-critic here points out that the claimant (i.e., Holocaust affirmer) has a Jewish grandfather that was killed by the Nazis in World War Two, and the fact that the Nazis killed his grandfather is the real basing reason that the claimant uses to support the Holocaust reality. The decisive evidence he has for the conclusion isn't the high-quality historical research; the decisive evidence is the death of his grandfather at the hands of the Nazis. Certainly, for Kelly's burden of judgment to succeed, the person's privileged self-knowledge claim to particular basing reasons must actually be about those particular basing reasons.

It might seem odd that someone would be mistaken about what their basing reasons are. But, in Part One I show how Carruthers cites good empirical studies that show people don't often know the real basing reasons they have for deciding an issue (Carruthers, 2011). Consider again the empirical research of (Valins, 1966)

showing that people often believe they are attracted to an image of a woman based on the attraction felt, but they actually believe based on a false belief that their heart was beating faster. Consider again (Bem 1970) showing that people often believe that students shouldn't have control of what is in their curriculum because they thought this through thoroughly, but in fact they believe simply because they didn't get paid to write an essay arguing for no such control. Or consider another study we talked about where people believe the socks on the right side of the aisle are higher quality based on their examination, but rather their judgment of quality was solely based on their conventional expectation that socks displayed on the right side are better. Or consider one of our four paradigm target disagreements given in the introduction, JEALOUS**;** here the person believes she hates the coworker based on the person being mean, but she really hates the person because she is jealous. People often don't directly know the basis upon which they make a decision or evaluation.

We can see another example if we extend Kelly's juror case study with the person who thinks the defendant is guilty pushing back against the protagonist who has the privileged self-knowledge claim that the basing reasons for her view is the evidence that both jurors share in common. Were the juror that argues for a guilty sentence a strong peer he would likely be able to point out challenging evidence that the juror who argues for innocence doesn't have the basing reasons he thinks he has. Say the claim-critic here points out based on the observation of the claimant behavior and speech that the claimant identifies with the defendant's politics and business acumen, and the claim-critic asserts that the claimant bases her assessment on politics and business acumen rather than the evidence they share together.

The problem with Kelly's burden of judgment approach to peer disagreement is that it doesn't express a way of accounting for the fact that people often can't reliably pick out their actual basing reasons for a judgement. Privileged self-knowledge about basing reasons is required for the burden of judgment approach Kelly argues for. The major criterion that Kelly expresses for satisfying the burden of judgment approach is that one has basing reasons that entail the conclusion one has in a peer disagreement. If you have compelling basing reasons in your initial reasoning, then it is rational to stay with your initial judgment without having to perform the post-initial-thinking reliability assessment Christensen's Independence Thesis demands. Yet, even if one's basing reasons closely entail the conclusion, if one doesn't really use those basing reasons one claims

to have, then the burden of judgment approach doesn't work. Kelly could fix this by adding another criterion, namely, that the person indeed does have the effective basing reasons one thinks one has. But that seems more like a bandage. What we really want is a remedy that tells us when the privileged self-knowledge claim likely can be trusted and when it can't, a remedy that doesn't just add more stipulations to what the burden of judgment means.

And we have indications of a remedy in the research and case study JEALOUS. We have found that privileged self-knowledge claims are not invulnerable, since they are able to be indirectly scrutinized by observing the behaviors and speech that the claimant expresses. In the examples of strong peer disagreements above we have shown how observation of behavior and speech can imply a person doesn't have the privileged self-knowledge they claim to have. We will talk much more about this in the next chapter, but here we can present the technical way of describing this:

> **Indirect Scrutability Norm**
> If one disputant has observational-interpretive evidence that implies the other's privileged belief about the self is unjustified, and shares the observational-interpretive evidence with the other, then the other's justification for this privileged belief is defeated.

The Indirect Scrutability Norm will be shown to be integral to the final remedy expressed in the last chapter, the Assessment Framework. The Assessment Framework will not just add criteria to our understanding of judgments. It will give us concrete signs to look for in order to gauge whether to trust or not trust privileged self-knowledge claims in peer disagreements about such claims.

Argument for widespread inadequate awareness of the fragile nature of privileged claims

We have uncovered a variety of crucial ways in which prominent scholars use privileged claims to define their own views and to respond to central issues in the literature while oblivious to the fragile status of privileged claims, and the following is a roundup:

Used as higher-order evidence confirming first-order evidence (Christensen):
1) "I know that I am paying attention."

Used to declare privileged self-knowledge of a particular mental state (Sosa):
2) "I know that I have a headache".

Used as higher-order personal information confirming first-order evidence (Lackey):
3) "I have decisive personal information."

Used as primary evidence for a belief (Michael Bergmann)
4) "I am extremely confident that p based on 'I am extremely confident in my insight about the way things really

are which implies p.'"

Makes rational deliberation possible (Kelly)

5) "I know my base reasons for my belief and I know they entail the conclusion."

While just surveying the work of five prominent scholars weighing in on central issues in the literature, we have found five different and pivotal ways of using privileged claims in a way oblivious to the fragile prima facie status of such claims. Let's go over what we know from this roundup.

In two of the five prominent scholars we see the least recognition of the fragile nature of privileged claims. As discussed, Sosa focuses on an atypical privileged claim epistemicly secure in peer disagreements as his main example of hidden but effective evidence, and in so doing avoids the more fragile nature of such claims especially in peer disagreements. He also gives the impression that one can't be wrong about "I have a headache," when we have demonstrated how one can. Bergmann denies any fragile nature in privileged self-knowledge claims to have personal extreme confidence in the trueness of foundational beliefs that, for example, perceptions appropriately indicate objects in the world. Such privileged claims are just the opposite of fragile, Bergmann thinks; they are the strongest way to hold one firm to one's beliefs in peer disagreement even with the highest-level of symmetry of internal and external rationality, as we will see in Chapter Five.

For two of the five prominent scholars, Lackey and Christensen, privileged self-knowledge claims about higher-order evidence are key to being able to hold on to one's belief at the same credence level in a peer disagreement. Yet they never adequately recognize the fragile nature of the prima facie true status of those claims. Without the recognition of the fragility of such claims to privileged self-knowledge some of the most challenging and complex peer disagreements can't even be recognized, since strong epistemic peers often challenge the higher-order evidence based on the fragile nature of the prima facie true status of those claims.

All five of the privileged claims we have found play crucial roles in the views of peer disagreement that prominent scholars develop in responses to key issues:

- Hidden evidence decisive for remaining steadfast (Sosa),
- Higher-order evidence that helps one decide who is more likely to get the issue right (Christensen),
- Personal information decisive for helping one stay steadfast (Lackey),
- The extreme confidence that allows one to stay steadfast in the most challenging peer disagreements because it is positively correlated with the reliability of insights (Bergmann),
- The ability to deliberate rationally knowing one's base reasons (Kelly).

We can now say there is widespread use of crucial privileged claims playing key roles in the development of the views and positions of prominent scholars towards peer disagreement and in response to key issues in this field. Given that very diverse and prominent scholars have demonstrated that privileged self-knowledge claims play crucial roles in their views of peer disagreement and responses to central issues in the field, given that many other scholars are highly influenced by the views and responses of these prominent scholars, it seems safe to say that such claims play pivotal roles widely in the literature and in the work of scholars in the literature. Because of the diversity of these prominent scholars, we can say that our sampling is representative of the diversity of scholars in the literature.

And we have seen that privileged self-knowledge claims have played pivotal roles in major issues in the literature:

- The difference between first-order and higher-order evidence and when it is appropriate to use each in peer disagreements,
- Internalist versus externalist views of justification in peer disagreements,
- The importance of basing relations,
- The epistemic efficacy of hidden evidence,
- The reliability of high confidence
- What constitutes peerhood.

And the list could go on. While we have not gone into deep details about some of these issues, the overlay has helped us see the importance of the fragile nature of privileged self-knowledge claims for understanding these issues. It seems safe to say that privileged self-knowledge claims play crucial roles in how major issues in the literature are resolved.

Likewise, it seems safe to say that privileged self-knowledge claims play crucial roles in the various positions in the peer disagreement literature:

- Steadfastists,
- Conciliationist,
- Independence Thesis,
- Total Evidence,
- Justificationist,
- Right Reasons
- Externalist about justification,
- Internalist about justification.

The overlay we engaged is not perfect. It didn't yield a complete and comprehensive understanding of all the major issues, the major positions, all the roles of privileged claims, all the scholars of peer disagreement,

and the complete varieties of scholars. But it does give us the ability to reasonably claim that privileged claims are crucially important widely in the literature on peer disagreement.

There is one more summary of our findings that we must give before it is appropriate to present the conclusion of this chapter: A summary of the many advantages that a more balanced understanding of the fact that privileged self-knowledge claims appropriately have prima facie true status, on one hand, and the fact that the prima facie true status is fragile, on the other hand. Right now they aren't balanced. The literature right now mostly honors the fact that privileged self-knowledge claims have prima facie true status, and the literature is largely oblivious to the fragile nature of the prima facie true status of sincere claims to self-knowledge based on privileged access. We need to honor the fragile nature of the prima facie true status <u>just as much</u> as we honor the prima facie true status itself. Let me try to move in that direction by giving a summary of the advantages we described that we will have if we honor the fragile nature of the prima facie true status as much as we honor the prima facie true status itself:

Benefits if take fragile nature seriously

- First, …. A more accurate understanding of peer disagreement as it happens in everyday life.
- Second, …. More realistic understandings and case studies presenting stronger peers.
- Third, …. A more accurate understanding of the complexity of peer disagreements.

First, if we correct the imbalance, we will have a more accurate understanding of peer disagreement as it happens in everyday life. In the everyday trenches of peer disagreement, people call into question the prima facie true status of their privileged self-knowledge claims, whether they be about privileged self-knowledge claims that give use privileged insights and information about one's grounding reasons for a conclusion, about the reliability of one's higher-order evidence, about the evidence one feels is hidden but still effective, about one's extreme confidence, or even about the personal information that one thinks one has about oneself like whether I am paying attention. Here are a variety of diverse examples of everyday challenges to the prima facie true status of privileged claims in peer disagreements:

- "I know you sincerely think you love me, but you really don't. Your behaviors to me clearly show you don't love me"
- "You say you aren't jealous of your brother, and I think you really believe this, but it is so clear that you really are."
- "You tell yourself that you love the study of philosophy, but really you just want the approval of your mother who is a famous philosopher."
- "While you say you are extremely confident that this new invention will make you millions of dollars, and that it just seems so incredibly clear that it can't be wrong, how can anyone be so confident?"

- "I know you have unimaginable confidence that Krishna appeared to you telling you to be a sannyasin, as clear as I am standing in front of you right now, as clear, you say, as your sense data of a hard surface with four legs protruding downward to the floor can be interpreted to be a chair; but couldn't it just be because you psychologically so badly wanted to be visited by Krishna?"

Maybe you noticed that each one of these examples of everyday peer disagreements exemplifies one of the diverse uses described above of how privileged self-knowledge claims are deployed in the literature in peer disagreement. Peer disagreements using privileged self-knowledge claims to support one's case and peer disagreements where such claims are what the disagreement is all about frequently happen in everyday life. It can be uncomfortable to challenge a belief that someone has about themselves, and also to be on the receiving end, but we often have to do this. Why, because we know that people often make mistakes about these things, because we know that people often have ulterior motives influencing what they think they know about their own mental states, and because we know that preferences, cultural views, psychological issues, ambitions often influence privileged self-knowledge claims. We challenge privileged claims in everyday life because we intuitively know that their trueness is as appropriately taken prima facie as is their prima facie status appropriately taken as fragile. This is the balance that normally takes place in everyday life, and the literature on peer disagreement will only reflect everyday peer disagreements when it can match this balance.

Second, if we correct the balance there will be more realistic understandings and case studies presenting stronger peers. While the case studies of scholars we have discussed claim to represent disagreements in everyday life, they often don't represent the fragile nature of privileged claims that occurs in everyday life. If the case studies that we have referred to in this chapter presented by prominent scholars of peer disagreement are typical of case studies in general in the literature, and I would argue that they are, then the literature doesn't present strong peers in general. We saw Christensen use, without any pushback from the peer, a privileged claim about higher-order evidence of attention to support a protagonist staying steadfast. I want to know how Christensen's Independence Thesis can handle cases where a strong epistemic peer pushes back on this privileged claim to higher-order evidence with reasonable evidence of fragility that calls into question legitimately the prima facie true status, especially since his Independence Thesis so heavily depends on having higher-order evidence that unambiguously indicates the initial thinking into whether the peer disagreement was either good or bad. Sosa doesn't consider a strong peer's challenge to his hidden-but-effective peer

disagreement view that he supports with the example of people who know there are stars in the sky with the evidence for that claim hidden yet effective. A strong peer here would likely point out the atypical nature of the paradigm case used to support this view, atypical because it doesn't reflect the stronger fragility of instance where hidden-but-effective evidence is used, as we described above. Bergmann hinges the justification of moral claims, religious beliefs, and beliefs about the legitimacy of claiming we can know objects based on perception on the privileged claims to extreme confidence given the seeming of "felt veridicality". I want to see how well his arguments for moral principles and beliefs in God during a peer disagreement work when there is a stronger peer pushing back by asserting as fragile the idea that one can use the same strategy responding to all these issues with the same "extreme confidence" given the felt veridicality. I also want to see how Bergmann's steadfastists view can be defended when a strong peer points out all the empirical research that concludes privileged "extreme confidence" used as a basis for a belief is highly fragile. As scholars recognize more appropriately the fragility of the prima facie true status, they will incorporate into their case studies and also into their peer disagreement theories stronger peers.

Third, we gain a more accurate understanding of the complexity of peer disagreements, if we encourage a more balanced understanding of the prima facie true status when we also insist on factoring in the fragility of that status. Christensen, Sosa, Lackey, Bergmann, and Kelly, and other scholars of peer disagreement may have fantastic ways of taking into account the fragile nature of the prima facie true status of privileged self-knowledge claims, and have excellent ways of presenting and defending against stronger peer characters in their case studies. And when they do their theories of peer disagreement will be much more complex, since their theories will need to account for all the many crucial ways that privileged self-knowledge claims are used in many aspects of the epistemology of peer disagreement. They will have stress tested their theories and case studies as a result. And this will undoubtedly result in a more complex understanding of peer disagreements, and this will likely match the complex ways in which peer disagreements are engaged in everyday life.

Fourth, we gain a more accurate understanding of the complexity of peer disagreements, if we recognize the interrelatedness of the observational-interpretive and the privileged access methods of knowing oneself. We have seen that Christensen and Bergmann explicitly think the two methods are segregated. The other

scholars treated here implicitly go along with the Inscrutability Thesis. Often in target disagreements in everyday life the claim-critic challenges the claimant with observational-interpretive evidence, and the Inscrutability Thesis doesn't have a way to adequately account for the challenge based on observational-interpretive evidence. In Part Three of this work, we will see that the Assessment Thesis remedies this oversight.

Now we can get to the conclusion of our Chapter: Privileged self-knowledge claims play essential and crucial roles in the epistemology of peer disagreement. They have key positions in the epistemologies of scholars in the field. Their prima facie true status is fragile in ways we have described. We have seen the many benefits of a more balanced treatment of the prima facie true status, on one side, and the fragile nature of that status, on the other. The literature is largely oblivious to the fragility of privileged self-knowledge claims. There are clearly many benefits to a more balanced treatment for the literature as a whole. Now we understand the peril and risk if we don't rebalance the prima facie true status of privileged claims by adequately taking account of the fragile nature of this status.

Chapter Five

The Scrutability Thesis Taken to the Extreme and the Indirect Scrutability Norm as Response

This chapter describes two epistemic norms that serve as guidelines for how far we can trust any privileged self-knowledge claims made in peer disagreements. I develop these norms because we have seen in Chapter Four that even some of the most prominent scholars of peer disagreement (and indeed scholars of peer disagreement in general) have made unwarranted assessments of epistemic peer status because they have not critically understood how far one can trust privileged self-knowledge claims in peer disagreements, and because these two epistemic norms show us just how to fix that problem. We begin first by describing the two norms in a preliminary way, and by describing why we can only have the best understanding of them in the process of evaluating Michael Bergmann's unique way of using privileged self-knowledge claims in peer disagreement. We then describe Bergmann's maximally symmetric peer disagreement set up with full disclosure which ends in full rationality for both disputants' mutually-contradicting beliefs; and this disagreement also ends with both disputants rationally remaining steadfast in their own respective privileged self-knowledge claims which led them to their respective mutually-contradicting beliefs. Next, we show how what we will call the Indirect Scrutability Norm appropriately diagnoses and remedies the oversight seen in the literature due to the Inscrutability Thesis. At this point we will be able to describe how the two are norms not just for scholars of peer disagreement but indeed for anyone in everyday situations of peer disagreement. For addressing many of life's most important questions, Bergmann has the same strategy: Demonstrate how a particular belief is implied by a privileged self-knowledge claim to an insight that implies that belief. He uses this strategy to answer the following questions: Is the skeptic right that there is no knowledge? Is there a God? What is the right morality? And you might think this strategy is unpopular. But there is no denying that it is widespread and extremely consequential in the public domain. Consider just one example: On the basis of his privileged self-knowledge claim that Jesus was present to him helping him win a crucial battle in 312 CE, Constantine believed all his subjects should be Christian. Other more secular examples will be given in business,

psychology, and interpersonal relationship. Indeed, the two norms help gauge how far anyone can trust privileged self-knowledge claims however consequential the peer disagreement.

This chapter develops five of the six key components of target disagreements that are taken into consideration when deriving the Assessment Framework. It discusses the fragility of privileged self-knowledge claims and the two methods of gaining self-knowledge. It develops both the Prima Facie Norm and the Indirect Scrutability Norm. It develops these norms in terms of the work of Michael Bergmann. This chapter shows the fragile nature of privileged self-knowledge claims by pointing out a mistake about the nature of such claims from one of the best philosophers who discusses peer disagreements. Bergmann's work on peer disagreement is an example of how even the most sophisticated intellectuals both don't take adequately into account the fragile nature of privileged self-knowledge claims and deny the interrelatedness of the two ways of knowing oneself imbedded in the Indirect Scrutability Norm.

The two universal epistemic norms and why we develop them in terms of Bergmann's views

Many scholars and people in everyday life have the intuition that a privileged self-knowledge claim is prima facie likely true when stated sincerely, for example, "I was paying attention." And many also have the intuition that high confidence increases even more the likelihood that it is true, for example, "I am highly confident that I was paying attention." Empirical research on self-knowledge and mindfulness described in the First Part confirms the appropriateness of these widespread intuitions. The Prima Facie Norm stated below formalizes the appropriateness of normally trusting these intuitions. Yet there are always limits to this trust. The Indirect Scrutability Norm stated below specifies one condition in which those intuitions should not be trusted in a prima facie way. The Indirect Scrutability Norm is developed in this chapter in the process of defeating the Inscrutability Thesis discussed in the last two chapters, which, as we have seen, says the privileged self-knowledge claim of a person can't be scrutinized by another person. The Inscrutability Thesis is crucial for Christensen's, Lackey's, Sosa's, and Bergmann's responses to peer disagreement, as we saw; yet we will see how it is false because it doesn't adequately acknowledge the fragile nature of privileged self-knowledge claims, and because it doesn't adequately acknowledge the integration of the two methods of gaining self-

knowledge. We also saw in Chapter Four that many prominent scholars have not adequately factored into their understanding of peer disagreement the limits of intuitions about privileged self-knowledge, and we saw how this often significantly leads them to make unrealistic assessment of the disputants' epistemic positions (like in Michael Bergmann, 2009; David Christensen, 2018, Jennifer Lackey, 2010). This chapter works towards correcting this lacuna in the literature for a more realistic understanding of peer disagreements by developing the two following epistemic norms telling us how far we can trust these intuitions about privileged claims:

> **Prima Facie Norm:**
> Trust a disputant's privileged self-knowledge claim prima facie if there are no or little signs that it is unjustified, and adjust credence levels of the claim according to the following scale: No signs = highest credence; little signs = high credence; significant signs = low credence; highest signs = lowest credence.
>
> **Indirect Scrutability Norm:**
> If one disputant has significant or highly significant observational evidence that implies the other's privileged self-knowledge claims is questionable, then don't trust that claim according to the following credence levels: highest signs = lowest credence; significant signs = low credence; little signs = high credence; no signs = highest credence.

These two norms give us a "floor" in the Prima Facie Norm and a "ceiling" in the Indirect Scrutability Norm for understanding how far in <u>any</u> peer disagreement we can trust the common intuition that people are accurately describing their mental state when they sincerely assert a privileged self-knowledge claim. The "any" is underlined in the last sentence for good reason.

And let's just review the different kinds of privileged self-knowledge claims that we have seen so far in peer disagreements, and we do this in order to assure ourselves that the "floor" and "ceiling" truly apply to <u>all</u> peer disagreement that involve privileged self-knowledge claims. In this dissertation we have described many of the functions that privileged self-knowledge claims play within a peer disagreement. In Chapter Four we saw them function as higher-order evidence (Lackey, Christensen), as primary evidence for a belief (Bergmann), as basing reasons for a belief (Kelly), as a declaration of hidden but effective evidence (Sosa). The four case studies listed in the Introduction to this dissertation use privileged self-knowledge claims to establish their beliefs in similar ways. In this dissertation we are claiming that the two norms developed here apply to all these different functions of privileged self-knowledge claims in peer disagreements. In each one of the peer disagreements listed in the Introduction, the privileged self-knowledge claims are initially taken as having prima facie warranted status (see the Prima Facie Norm). And in each they are challenged because observational-interpretive evidence isn't in sync with the mental state the privileged self-knowledge claim

points to (see the Indirect Scrutability Norm). And each of the five privileged self-knowledge claims in Chapter Four is challenged based on the fact that the observational-interpretive evidence isn't in line with what should be expected, if the privileged self-knowledge claim reflected reality (see the Indirect Scrutability Norm). The two norms apply to <u>all</u> situations of privileged self-knowledge claims in peer disagreement. That is a remarkable find, and this will, as said in Chapter Four, help scholarship reflect more the reality of peer disagreements.

There are two beneficial ways of categorizing the different types of privileged self-knowledge claims. They will help us understand better the nature and application of the two norms, and they will illustrate why it is crucially important to develop the two norms in terms of a treatment of Michael Bergmann's understanding of peer disagreements. A first classification will be helpful between those privileged self-knowledge claims that apply to deep issues of life—like morality, philosophy, religion, and self-concept—and those that are not deep. A second useful classification divides deep privileged self-knowledge claims into two groups, the ones that focus on the privileged self-knowledge claim itself as the target of the peer disagreement, and the ones that don't focus on the privileged self-knowledge claims itself. For example, Bergmann's key privileged self-knowledge claim under "Not focus" in the illustration below plays a supporting role for justifying p, and the focus of the use of the privileged claim is p. Among the five privileged self-knowledge claims focused on in Chapter Four and the four presented in the Introduction, there are three centrally focused on the privileged self-knowledge claim itself, and there is one that doesn't present the privileged self-knowledge claim as the central focus:

Deep

Focus

"I know I love you!"

"I know I am not jealous."

Not focus

"I am extremely confident that p based on 'I am extremely confident that my insight about the way things really are includes p.'" (Bergmann)

As you can see from the illustration just above, there are two privileged self-knowledge claims that are deep and also are the central focus of their respective peer disagreements that are described in the introduction. While it is obvious in their case studies that they are the central focus of the peer disagreement, the fact that

they are deep isn't so obvious. "I know I love you!" and "I know I am not jealous" are considered deep here because they have huge implications for one's self-concept; and when such claims have implications for one's self-concept, there can be more unconscious ulterior motives present blocking self-knowledge. Of course, any issue can have implications for one's self-concept if one is personally highly invested in it, but these seem particularly vulnerable.

One of the best ways to develop these two epistemic norms is to show how they properly diagnose and correct what exactly is wrong with Michael Bergmann's sophisticated, but ultimately mistaken, argument. He argues that one can stay steadfast in any peer disagreement with a belief supported by a privileged self-knowledge claim casting off any challenge so long as one maintains the grounding privileged self-knowledge claim. He presents one of the most sophisticated arguments for a moral principle based only on a privileged self-knowledge claim, and we will see how well the norms hold up when tasked to evaluate Bergmann's argument.

Bergmann uses privileged self-knowledge claims in a sophisticated way as evidence for some of the most consequential beliefs, such as the belief that one can appropriately conclude there are objects through perception, that we are in a real world as opposed to a matrix, that God exists, and that it is immoral in principle to randomly kill people. Of all the scholars of peer disagreement we have looked at, only Bergmann uses the extreme confidence in a privileged self-knowledge claim as the most central grounding of his entire position on peer disagreement. You may not agree with his use of a privileged self-knowledge claim as an epistemic foundation supporting a view in a peer disagreement. But such strategies are very popular and often extremely consequential in religion, ethics, and politics. Witness David Koresh's privileged self-knowledge claim to be the second coming of Christ, or Constantine's belief that all his subjects should be Christian based on the privileged self-knowledge claim that he saw a definitive sign in the sky demanding this.

Bergmann's maximum symmetry set up with steadfastness on both sides

We begin first by describing a few key concepts that Bergmann uses, Alvin Plantinga's distinction between internal and external rationality, insight, and error theory. An insight is an experience that precipitates a particular belief. Insights are mental states, and, so, as we have been saying, and as Bergmann also thinks, one has first person authority to them when claiming them based on privileged access. You can disclose to another your insight and mental state, but the other can't have it. One disputant, S_1, in the disagreement affirms p based on an insight, while the second disputant, S_2, affirms ~p based on her insight. Each has their own error theory about how the other's insight isn't genuine and how the propositional content of the beliefs formed on the basis of the insights are false, where "error theory" is a view about how the other's attempt at affirming true beliefs went wrong. Finally, Bergmann describes internal rationality as the epistemic belief formation that occurs after, and in response to, the subject's experience of a mental state like an insight. External rationality is the result of cognitive processing mechanisms working in an environment to produce experiences such as an insight.

Now we can describe Bergmann's theoretical setup of a peer disagreement with full disclosure NS perfectly symmetric evidence on both sides—that is, again, except each disputant's privileged and mutually-contradicting self-knowledge claims. Many scholars think it is impossible, or nearly so, to disclose all one's relevant evidence (King, 2012; Sosa, 2010). We will describe how both disputants can be steadfast in this setup context. In this theoretical set up, both disputants are epistemic peers with equally rational and mutually-contradicting privileged claims held with extreme confidence, even while disclosing all evidence on both sides. Each disputant is internally and externally rational while each stays confident in the epistemic adequacy of their own belief. Thus, each can see the other as externally and rationally deploying their cognitive processing mechanisms. At the same time each deploys their error theory to show how the other has a mistaken belief in response to their experience due to no epistemic fault of their own, since the other's environmental content was abnormal in some way.

For example, say Breanna and a friend are both looking out the same window, and Breanna tells her friend she loved seeing the beautiful bird through the window that just flew by. Her friend is shocked saying there was no bird that flew by the window just then. Both can have externally and internally rational responses to

their respective experiences, if they both formed rationally appropriate conflicting beliefs appropriately responding to the different experiences, and if they rationally formed further beliefs from the one's generated by the consequential experience. Say Breanna is wrong, it wasn't a bird flying by; it rather was a floater that just formed on her retina, which commonly appears as one gets older. Even though Breanna falsely formed a belief that it was a bird, that false belief was an externally rational response to her experience, given that floaters often look just like a bird in flight. And each thinks the other has the right internally rational response to the belief formed as a result of the experience. In this case there is epistemic peer disagreement with each disputant fully internally and fully externally rational after full disclosures on each side, even when each rationally thinks the other is wrong about the bird flight, and even though Breanna was wrong through no fault of her own.

Bergmann's description of how to stay just as confident

Bergmann realizes that the maximally symmetric peer disagreement can lead some people to have serious doubts about one's epistemic status in such a maximally symmetric peer disagreement, and this doubt can result in an undercutting defeater whereby one's basis for believing and being steadfast are discredited. For his paradigm example of someone whose evidence is being undercut he describes a woman who is having some doubts about the reliability of her belief based on her consequential experience, and we can give her a name though Bergmann doesn't, Asia. Asia concludes that the other disputant is in a maximal and epistemically symmetric position also with full externalist and internalist rationality, and she concludes "it's a live possibility for someone with roughly her degree of intellectual virtue to be highly confident in the ways just mentioned and yet be mistaken" (Bergmann, 2009, 344) through no fault of her own; and Asia consequentially questions her extreme confidence: Do I really have the privileged extreme confidence I think I have that my peer is wrong, given that my peer came to think I am the one mistaken through no rational fault of the peer's own, and given that right now I could just as easily be the one mistaken through no rational fault of my own? We can call this The Potentially Undermining Question:

The Potentially Undermining Question:

> Do I really have warrant for the privileged extreme confidence I claim for thinking my peer is wrong? My peer came to think I am the one mistaken through no rational fault of his own. I could just as easily be the one mistaken through no rational fault of my own.

Of course, Bergmann in the quote below answers Asia's question, no; and, since "Asia" is just a hypothetical person that refers to many people, this is Bergmann's answer as well to all the copious amounts of non-steadfastists who have reduced confidence in similar situations.

> I think that the rational response for S1 [i.e., the protagonist, i.e., Asia] to E [i.e., the details of the symmetry among peers] is to continue believing p. However, I don't have an argument for that conclusion, just as I don't have an argument for the conclusion that the rational response for us to a tactile experience like the one we typically have when grabbing a billiard ball is to believe something like "that's a small hard spherical object". I can see, in the case of the billiard ball, that that belief is a rational response to the tactile experience in question. But I don't have an argument for why that is so. Likewise, I don't have an argument for the view that the rational response for S1 to E is to continue believing p. Nevertheless, I will mention two examples that I think help us to see that this is so. (Bergmann, 2009, 345)

In the quote above Bergmann concedes that he doesn't have a knockdown argument for staying steadfast in this peer situation. Yet Bergmann thinks it is rational for someone like Asia to stay steadfast even though it goes beyond what one can rationally prove.

Temporary pivot away for more background

Now, we must temporarily pivot away from answering The Potentially Undermining Question knowing that we will come back to it with better background. And this temporary pivot is necessary because Bergmann in the quote just above gives us only his short answer to the question. To get Bergmann's longer and deeper answer to this question we must follow the reference Bergmann gives in the quote just above to the earlier work he has done on perception and objects because he uses the conclusion of this earlier work to justify the extreme confidence he talks about in peer disagreements. So, we will get back to this question once we have together all the background on Bergmann's thinking referenced in the quote that we need to give Bergmann's longer and deeper answer to this question.

In the quote just above, it seems contradictory to say that you can have rational retention of a belief without being able to rationally prove that belief. In the very next sentence Bergmann shows us where to get an explanation for what this means and how it can make sense. Bergmann refers us to his work on perception and skepticism in order to resolve the apparent contradiction and best understand his deeper answer to The Potential Undermining Question. It is reasonable to retain a belief in the completely symmetric peer disagreement even

without a rational argument just as it is reasonable to form beliefs about things and objects based on perceptions even though one can't give proof that this is reasonable.

Since these two situations are similar, let's look more into his work on perception and skepticism. He has written about this many places, but we can look at what he says in (Bergmann, 2017). Here he says that the first step (which he says is inspired by Thomas Reid and William Tolhurst) to seeing the rationality of forming beliefs based on sense perceptions even though you can't give a rational argument for this is to highlight the strong epistemic intuitions that one thinks the skeptic is wrong (Bergmann, 2017, 21). In the second step one finds that the epistemic intuitions one has for affirming the rationality of forming beliefs based on perceptions are stronger than the intuitions the skeptic has for denying it. We can find the final step in the following:

> Tolhurst calls this feel of a state whose content reveals how things really are its 'felt veridicality'. It is the distinguishing feature of seemings. …These higher-order seemings that our first-order beliefs are reliably formed or that our first-order seemings are veridical are examples of epistemic intuitions. …The Reidian reply is to reject those doubts and affirm the reliability of those beliefs and the veridicality of the seemings on which they're based. In doing this, the one endorsing the Reidian reply is forming a higher-level belief about the trustworthiness of our faculties and this higher-level belief is based on epistemic intuitions." (Bergmann, 2017, 22-23).

We find out that the non-skeptic epistemic intuitions are stronger by considering the higher-order evidence. The higher-order evidence comes to us as seemings, and seemings assure us that our first-order beliefs are reliably formed; they tell us "how things really are". On the basis that higher-order seemings indicating first-order beliefs from sense perceptions are reliable and veridical, we can reject the skeptical view.

Bergmann illustrates with a case study what he means when he says one can rationally retain a belief even when one doesn't have a rational argument for it. This case study of a peer disagreement also has symmetric, complete rationality on both sides, full steadfast stances on both sides, and fully disclosed evidence on both sides. The issue in the peer disagreement involves whether the actions of a serial killer, Jake, are morally wrong. Jake maliciously kills families randomly forcing parents to watch their children tortured to death. The protagonist has two friends he is in a disagreement with, one an ethical egoist, the other a moral nihilist, neither believing what Jake does is morally wrong in principle. All three find the actions of Jack horrendous and should be deterred by punishment. All three have externally rational beliefs about this matter that they are confident are true, each has internally rational beliefs based on their experience. Each has an adequate error theory explaining how those that don't agree with them are mistaken. Each thinks those who don't agree with them

have a wrong or misleading experience through no fault of their own, so that means they still have external rationality even when they have a mistaken intuition. Each sees the others as having the same epistemic virtues. The only asymmetries are their different privileged self-knowledge claims, their protagonist with the claim "I am extremely confident that my belief is genuine", and with the two friends who both claim "I am extremely confident that ~p," and apart from this there is complete symmetry. Each sees the others as completely rational, both internally and externally; each sees the others as mistaken through no fault of their own because the environment wasn't cooperating, that disputant was subjected to unlikely error despite reliability.

After following the lead that Bergmann gives us in the quote from the 2009 disagreement article we have been analyzing, we can now have a better understanding of what Bergmann means by rationally retaining a belief even though one can't give a rational argument for it. What he means is that we can't give a rational argument for the belief that Jake's actions are morally wrong, just like we can't give a rational argument that it is rational to form beliefs about objects based on perceptions. But, by considering the higher-order evidence of seemings that we have about the reliability of our first-order beliefs, we can rationally retain our first-order beliefs even without being able to argue a rational case for those reliably formed first-order beliefs. So, it appears that the reason that Bergmann has for thinking one can't give a rational argument either for the rationality of forming a belief on the basis of perceptions or for staying just as confident—when one finds out one has maximum symmetry with one's opponent except for respective asymmetric Privileged Claims and it seems one could have just as easily made the mistake one's opponent made equally through no fault of one's own—is that the seemings which lead both sides of the disagreement to have confidence that their belief is reliably formed has first person authority, and consequently one can't give them to somebody else as a proof or argument; one can only just tell somebody that one has those seemings and encourage others to try to get in an environment where they have the right seemings.

In light of this understanding by following Bergmann's referral, we can make better sense of the following crucial passage along with the contents of its footnote from the 2009 disagreement article we have been looking at.

> What makes this reaction plausible, I think, has something to do with the extremely high confidence we have that Jack's behavior is morally wrong, that our belief about the morality of Jack's behavior is reliably formed and that those who don't see that Jack's behavior is morally wrong lack some genuine insight had by the rest of us.[18] [see footnote [18] just below]…. The intuitive support for your own reliability on the topic of the morality of Jack's behavior significantly outweighs any reason to doubt that reliability that is provided by your recognition that someone equal to you in intellectual virtue could have your same level of confidence for parallel views (opposing yours) and yet be mistaken. (Bergmann, 2009, 347)
>
> [18]Tolhurst mentions something that helps to explain what's going on here. He speaks of a felt veridicality, which is a component of seemings. This felt veridicality is "the feel of truth, the feel of a state whose content reveals how things really are" (1998, 298–9). Apparent insights will produce seemings or intuitions, along with their accompanying felt veridicality. Tolhurst goes on to note that there can also be second-order seemings when one reflects on the felt veridicality component of a seeming: "When we become self-consciously aware of a seeming it seems to us that the seeming is veridical. This second-order seeming is grounded in our awareness of the feel of veridicality" (299). So when I speak, in the text, of a person's high confidence that p, we can think of the seeming on which that confident belief that p is based as involving a very strong feeling of veridicality. And the reason that high confidence that p tends to be accompanied by a high confidence that the belief that p is reliably formed is that, when one reflects on the strong felt veridicality of the seeming that p, it seems strongly to one that that seeming that p is veridical.

This passage says the belief that p ("Jack's behavior is morally wrong") is based on an insight of the way things really are which implies p. For Bergmann, an "insight" comes from one's cognitive capacities interacting with the environment (Bergmann, 2009, 341). Such an interaction produces experience which contains insights (Bergmann, 2009, 339). Insights produce seemings (Bergmann, 2009, 352), which are experiences of the way things really are (felt veridicality) (Bergmann, 2009, footnote 18). In those experiences of the way things really are, p can be an example of the way things really are (Bergmann, 2009, footnote 5). In (Bergmann, 2017, 22-23) we saw that he paraphrases Tolhurst's definition of felt veridicality as a state whose content reveals how things really are. And in the footnote quoted above he indicates this is a quote from Tolhurst. It appears this is also the way that Bergmann thinks about felt veridicality.

We have located the key privileged self-knowledge claim that allows one to remain steadfast in a maximumly symmetric peer disagreement about a belief whether that belief be an ethical one that a serial killer's actions are wrong morally, a philosophical belief that one can rationally form beliefs based on perceptions, a religious belief that Muhammad received revelation from Allah through the angel Gabriel on Mount Hira, or a political belief that Trump was an excellent president:

> **Privileged self-knowledge claim used as evidence to break a tie in an otherwise completely symmetric peer disagreement (Bergmann)**
>
> > "I am extremely confident that p based on 'I am extremely confident in my insight about the way things really are which implies p.'"

In the formulation above, there is both an assertion of p and a privileged self-knowledge claim used as a basing reason for the former. The two are formulated as independent claims (that is, "I am extremely

confident that p" and "I am extremely confident in my insight about the way things really are which implies p"), because Bergmann thinks of them as such.

We must take caution with Bergmann's use of "extreme confidence" in this argument for the steadfast view. The argument is about p, not about the "extreme confidence" that p. Bergmann seems to be adding this phrase "extreme confidence" in order to claim more justification for the belief that p. But we should focus on how well the privileged self-knowledge claim supports p rather than on the "extreme confidence." Also, extreme confidence in eyewitness testimony, including during a court proceeding, has been shown by many empirical studies to be not very correlated positively with the accuracy of the eyewitness testimony; (Roediger et al., 2012) describes the many empirical studies.

Getting back to the Potentially Undermining Question and Bergmann's response

Now that we have a deeper understanding of Bergmann's views, we can understand his response to people who start to question their extreme confidence in light of the most challenging disagreements (Bergman, 2009, 344). Here again is the best formulation of this worry:

The Potentially Undermining Question:

> Do I really have warrant for the privileged extreme confidence I claim for thinking my peer is wrong? My peer came to think I am the one mistaken through no rational fault of his own. I could just as easily be the one mistaken through no rational fault of my own.

Asia, our paradigm character representing the many people who start to have such doubts, begins to think that she is just as likely to have made the mistake through no fault of her own in the maximally symmetric peer disagreement. Bergmann says she can even in this situation rationally retain her confidence and credence level, but she can't be given a rational argument to this effect. And that is because, as we have seen, the only evidence she could have is from privileged self-knowledge claims to extreme confidence in response to her belief that she sees the way things really are. That experience can't be given to anyone, people have to actually have the experience in order to use it as evidence. Nobody can give a rational argument because nobody can give another the experience of the way things really are and the belief based on it.

Bergmann's recommendation to the person having The Potential Undermining Question, represented by Asia, is to get back in tune with the experience of the way things really are, felt veridicality. Foster the memory

of it or try to get the experience again. If you are not necessarily sure you have even had it, try to have it for the first time. Since every day we experience the felt veridicality that it is appropriate to infer objects from perception, one can experience every day the way other things really are, for example, the reality of Jack's behavior being morally wrong.

Bergmann also recommends that you remind yourself of the good higher-order evidence you have of your belief based on the seeming experience of the way things really are. The extreme confidence one has in the genuineness of one's seeming experience that supports one's belief that p is higher-order evidence that the seeming experience is actually genuine. In the 2017 article way of putting it, her high confidence is the result of higher-order seemings that her first-order beliefs are reliably formed. So long as you hold onto "a high degree of confidence in the genuineness of your own insights on this matter" (Bergmann, 2009, 347, see also 349), she has evidence for the reliable formation of the belief that p, she has evidence that she is vastly less likely to have made a mistake through no fault of her own, and she doesn't have such evidence for her opponents view that ~p. Recall that what Bergmann means by "insight" is the initial cognitive response that one has to the seeming experience. The insight response thus represents the interaction of our cognitive faculties to the things as they really are. The extreme confidence one has in the genuineness of both the seeming experience and the insights gained from the seeming experience is higher-order evidence that the seeming and insights are externally rational and are actually true. Since the belief in p (for example, Jack's behavior is morally wrong) is rationally based on the truth of the seeming experience of the world as it really is, and since there is extreme confidence that this seeming experience is genuine and "strong" (Bergmann, 2009, 349), the belief in p also acquires extreme confidence.

Just to recapitulate the argument Bergmann presents so far for thinking it is rational to stay firm in one's belief that p in a maximum symmetric peer disagreement, let's review what we have before seeing the final part of the argument. Asia, who represents anyone who is having similar doubts, has been thinking that her epistemic peer—Blake—has external and internal rationality just like her, and he is just as cognitively virtuous as her. He just was unlucky to have an environment not conducive to true insight, and it wasn't even any cognitive fault of his own. Asia comes to think she might just as easily have made the mistake through no fault of her

own. Perhaps the insight she gained as the result of her cognitive abilities interacting with the way things really are is false through no fault of her own; the environment perhaps was abnormal leading her cognitive faculties, though working properly, to produce false insights. Even the strongest inductive argument with true premises, cogency, and a high probability of producing the truth can have a false conclusion if in a highly unusual environment. She could just as easily have been the one who had the unlucky environment and false beliefs based on it, with false higher-order evidence that her insights were genuine.

One thing Bergmann would recommend is to remind her what epistemic situation she was in before these doubts: She believes with high confidence that she has a genuine insight with a normal environment and cognitive faculties adequate for the task thus giving her external rationality. She believes with high confidence that her belief that p was formed through a reliable process from its rational basis of the insight thus giving her internal rationality. And with the extreme confidence in the genuineness of the insight from the seeming experience of the way the world really is, she has even more evidence which is higher-order evidence that her belief that p is right.

The Inscrutability Thesis and indefeasibility

Inscrutability and the best response to the undercutting defeater

But, there is one more part of the argument needed for assuring Asia that she should stay confident. She needs a reason to favor her own evidence over the other's. Blake reports just as much evidence as she reports to him through full disclosure. He is just as cognitively capable, he is just as extremely confident that his insight conflicting with hers is externally rational, and he is just as extremely confident that his belief that ~p is correctly formed as a response to the insight making his belief internally rational, and Blake reports he has just as much higher-order evidence as she does.

At this point Bergmann recommends that Asia take comfort in the exclusive evidence she has which can't be disclosed because the two respective alternative and privileged self-knowledge claims to extreme confidence come from privileged first-person authority. Each is extremely confident in their insights from seemings, each is extremely confident in the beliefs they formed about p, and each is just as confident in the higher-order

evidence. But, neither can find out if the other really has the extreme confidence, neither can know the other's quality and quantity of confidence and evidence. You aren't even able to compare the strengths of the insights on both sides. You don't even know if the opponent has evidence and responses exactly parallel. Since you know that your insights are strong and you would be surprised to learn the other has equally strong insights, it is reasonable to conclude yours are stronger. In making this last move, Bergmann is assuming the very same Inscrutability Thesis we saw Christensen and others uses in the last chapter. We formulated the Inscrutability Thesis from Christensen's work:

> **Inscrutability Thesis** (a paraphrase of what is said in Christensen, 2011, 11)
> One can't know the privileged self-knowledge claims of others, and this fact is a good reason to take one's own evidence more seriously than the opponent's in a peer disagreement where every day, non-exotic day, non-exotic mistakes have been eliminated by multiple reliable methods, and where the only possible mistakes at play are exotic ones which each is able to rule out only for themselves, since one can know that one has ruled out the exotic mistakes at play and one can't know the same about the other disputant's exotic mistakes.

Bergmann has a less specific abbreviation of this Inscrutability Thesis with Bergmann's text that supports it just below:

> **Inscrutability Thesis Abbreviated**
> One can't know the privileged self-knowledge claims of others, and this fact is a good reason to take one's own evidence more seriously than the opponent's in a peer disagreement.

In this situation, Bergmann concludes, it is rational to go just with one's own evidence. You will see that Bergmann implicitly uses the Inscrutability Thesis as a crucial element in his steadfast position:

> As we noted earlier, they can't pass on these insights or their confidence in them to you. In fact they can't even really let you know what their precise confidence level is. Nor can they know what your precise confidence level is like. So it seems that none of you is able to compare, in an informed way, the strengths of the insights on both sides of this disagreement. But then it seems that you don't have very good reason for thinking that your ethical egoist and moral nihilist friends have evidence and responses that are exactly parallel to yours. And this might make you sensibly suspect that the apparent insights they have in support of the conclusion that it's false that Jack's behavior is wrong are (unbeknownst to them) significantly weaker than the apparent insights you have in support of the conclusion that Jack's behavior is wrong. (Bergmann, 2009, 349, see also 246)

> The intuitive support for your own reliability on the topic of the morality of Jack's behavior significantly outweighs any reason to doubt that reliability that is provided by your recognition that someone equal to you in intellectual virtue could have your same level of confidence for parallel views (opposing yours) and yet be mistaken. (Bergmann, 2009, 346 , see also 347)

> Given how strong your apparent insights are and how surprising it would be to learn that others (your equals in intellectual virtue) have equally strong apparent insights in support of an opposing view, the hypothesis that their apparent insights are (unbeknownst to them) significantly weaker than yours might seem to you to be the most plausible account of what's going on in this case of disagreement. (349)

Indefeasibility

Right now we have adequately described Bergmann's recommended response to people like Asia who worry that they could be the one mistaken through no fault of their own. But now let's press on to see how he thinks that a belief that p based on "I am extremely confident in my insight about the way things really are which implies p." defeasible. Bergmann sets it up such that the one who has extreme confidence can't have a defeater of their belief that p, if that person truly has extreme confidence. It is all about the right response to the mental state. Bergmann says there are only two ways that his privileged claim of extreme confidence can be defeated, if one doesn't believe it, or if one epistemically should not believe it:

> If in response to recognizing that S disagrees with you about p (which you believe), you either do or epistemically should [emphasis mine] disbelieve or seriously question or doubt the claim that you are, on this occasion, more trustworthy than S with respect to p, then your belief that p is defeated by this recognition; otherwise, not. (Bergmann, 2009, 343)

So, if one either does or epistemicly should doubt that one is presently more trustworthy for getting the issue right than one's opponent, then one's belief is defeated, otherwise, not. Notice Bergmann says he will in this paper only consider the epistemic "should" defeater.

Since one has to decide whether one is more trustworthy for getting the issue right, we now need to know when to doubt this trustworthiness. He then proceeds to reword the condition for defeat. He endorses the following more detailed conditions for doubting one's mental state and so defeating one's belief at issue in the peer disagreement:

> C: you are, on this occasion, more trustworthy than S (who disagrees with you about p) with respect to p;
>
> When *should* you disbelieve or seriously question or doubt C? When is that the epistemically appropriate response to recognizing that S disagrees with you about p? It depends on whether disbelieving or seriously questioning or doubting C is the epistemically appropriate response to your mental states – states which include your newly acquired recognition that S disagrees with you about p. (Bergmann, 2009, 343)

We can consolidate the statement earlier with the statement just above. So, if seriously doubting one's better trustworthiness for getting the issue right is the appropriate epistemic response to one's mental state, then one should seriously doubt one's belief.

So now we want to know when doubting or remaining steadfast is an appropriate response to one's mental state. The main question now: When is it an appropriate response to one's mental state to remain steadfast? In response to this question, Bergmann next surveys a number of epistemic views about what determines an

appropriate response to one's mental states (Bergmann, 2009, 343), like reliable indicators, whether one's belief fits one's evidence, in accord with proper functioning.

At this point, clearly frustrated by the lack of a clear consensus on how to evaluate responses to one's mental states, he says we have to look at specific cases to determine how to evaluate whether one's continued belief that p is an appropriate response to one's mental states (Bergmann, 2009, 343). So, he proceeds to do just that, that is, look at specific cases that illustrate when steadfastness is an appropriate response to one's mental states. He then shows how the undercutting defeater Asia is threatened by turns out not to be a problem for the reasons mentioned above. Disbelieving or seriously questioning one's better trustworthiness is not necessarily an appropriate response to the mental state one is in when having privileged self-knowledge claims to extreme confidence. If you hold on to your privileged self-knowledge claims with extreme confidence, then remaining steadfast is the best response to your mental states.

Given what we have described so far, when one has extreme confidence that p based on extreme confidence in one's insight about the way things really are which implies p, that person's extreme confidence that p can't be defeated. The only way that the extreme confidence in the belief that p can be defeated is if the person stops having extreme confidence in one's insight about the way things really are as including p. Any defeater can be turned away so long as one has extreme confidence in one's insight about the way things really are as including p. While it is one's insight about the way things really are as including p that justifies the belief that p, extreme confidence in the latter is what is needed to turn defeaters away in the most challenging peer disagreements like the one he presents. For Bergmann "extreme confidence" makes a difference. He explicitly highlights that it is extreme confidence that defeats all defeaters: "Suppose you have extremely high confidence in your apparent insight that Jack's behavior is morally wrong" (Bergmann, 2009, 349).

An epistemic norm about scrutability in peer disagreements

There are few things this concluding section proves before ending with a statement about how far one can trust intuitions about privileged self-knowledge claims in peer disagreements. First, we argue that the Inscrutability Thesis gleaned from Bergmann's and Christensen's texts is false; second, we establish that the

indirect scrutability of a disputant's privileged self-knowledge claim in a peer disagreement implies a possible defeater of it; third, Bergmann's and Christensen's theories of peer disagreement, as they now stand, can't account for (or gauge the epistemic status in) the types of disagreements we have given in the four examples of disagreements in the introduction; and this is a significant problem, since disagreements over privileged claims are some of the most important and consequential peer disagreements.

The Inscrutability Thesis is false

The Inscrutability Thesis gleaned from the texts of Christensen and Bergmann doesn't reflect how privileged self-knowledge claims are treated in everyday life (and in peer disagreements) and in the empirical studies on self-observation and mindfulness. In the last chapter we formulated the Inscrutability Thesis, and here it is stated again in its abbreviated form:

> **Inscrutability Thesis**
> One can't know the privileged self-knowledge claims of others, and this fact is a good reason to take one's own evidence more seriously than the opponent's in a peer disagreement.

The Inscrutability Thesis says one can't know another person's mental states specified in that person's privileged self-knowledge claim. The peer's mental states can't be inspected by others, whereas one can know one's own in a privileged way. And this assumption leads both of them to favor their own evidence over the opponent's in a peer disagreement. We can see how the Inscrutability Thesis is wrong, given what we have said in this chapter and the other two about the fragile nature of self-knowledge. This Thesis would work were it stated as applying only to direct knowledge of another's mental states referred to by a privileged self-knowledge claim.

But, we can have indirect knowledge of another's mental states pointed out by the person's privileged self-knowledge claims. This is what the research on self-observation and knowledge of others presented in Part One concludes. And furthermore, the empirical studies on self-observation imply that we sometimes do have self-knowledge indirectly through self-observation. Those studies show that we have self-knowledge of our mental states sometimes based on direct transparent access (see mindfulness studies detailed in Part One) and sometimes based on indirect observation. We can know another person's mental state indirectly because particular behaviors, speech, and body language are implied by a person's particular mental states.

In addition, our legal and mental health care systems depend on such indirect knowledge of mental states. To convict a person of a crime, you usually have to know the person had the mental state of intending the crime, that is, criminal intent. For a therapist to help a client with depression, usually she has to try to understand the client's mental states that are troubling. She can't get in the client's mind and transparently know his mental states with the kind of authority only the agent deciding an issue through deliberation has. She has to listen to the client, observe his behavior, tone of voice, body language, etc., in order to infer a mental state. Sometimes she will get her inference of a mental state from observation incorrect, but the success of her therapy largely depends on her mostly getting the inferences right. On the other hand, the client can know something about himself by taking seriously the inference of his therapist about his mental state which she achieved by observing him.

Certainly, we can't inspect the other person's privileged self-knowledge claim directly. That claim of the other is known to the other transparently when that person decides an issue, and that transparency is something that can only happen for the agent of the privileged claim. I don't know directly if a person is really paying attention when she says she believes she is. But I can scrutinize her mental state indirectly by observing her behaviors, speech, and body language.

Apart from all that evidence, there is evidence from peer disagreements in ordinary life that people sometimes know themselves and others based on observation. To the extent that the four examples in the introduction to the dissertation represent typical everyday peer disagreements where mental states are known through observation, to that extent they are evidence that such indirect ways of attaining knowledge of mental states are important and frequent. Each one of the case studies presented in the introduction to the dissertation shows how it is often appropriate to think that one can know the mental states of others based on their words, deeds, and speech, and this means that all of the four examples demonstrate how the Inscrutability Thesis is false. Privileged self-knowledge claims just aren't inscrutable. The four examples also demonstrate a possible defeater implied by the ability to know the other's mental state through observation.

Indirect Scrutability implies a defeater

The fact that there are two pathways for knowing the mental states of someone—one directly and another indirectly—opens up the possibility that there can be discrepancy between the two. For example, a wife accused of murdering her husband with a gun sincerely says, "I know it was an accident," and a juror thinks that she did intend to murder her husband based on the observation of the defendant and the physical evidence; the therapist telling the client he is jealous of a fellow employee based on the observation of her client's body language and speech, while the client himself sincerely thinks he doesn't like the workmate because he is selfish. The two different pathways—through observation, through privileged access—can be at odds with each other.

Given the two pathways to knowing one's mental states, a disagreement can ensue between two peers about whether one disputant's claim to self-knowledge based on privileged access has true propositional content. All the case studies in the Introduction are examples of such peer disagreements. In each example one disputant claims the other is wrong about the propositional content of the privileged self-knowledge claim at issue. The claimant says the claim's propositional content is true based on privileged access, while the claim-critic says it is false. In such a disagreement the claim-critic marshals all her observational-interpretive evidence that the claimant's alleged self-knowledge doesn't reflect the claimant's actual mental states.

The claim-critic argues that the observational-interpretive evidence of the claimant implies the claim to true propositional content is unjustified. Such arguments depend on the assumption that a person's behavior generally correlates positively with the person's state of mind. If I see a person act and talk in a way inconsistent with the person's privileged self-knowledge claim, this implies that the person's claim is unjustified.

The indirect knowledge of the other disputant's privileged self-knowledge claim in a peer disagreement implies a possible defeater of it. In each of the case studies in the Introduction, a disputant indirectly scrutinizes the other's privileged self-knowledge claim by observing the disputant's behavior and speech. And when one sees behaviors that go against what is implied by the mental states, this gives us justification to question whether the person claiming the mental state in a privileged way really does have that mental state. Those case studies are examples of peer disagreements where one appropriately defeats the opponent's justification by showing

how decisive observational-interpretive evidence implies the privileged claim is false. What we have said above leads to the following Indirect Scrutability Norm:

> **Indirect Scrutability Norm:**
> If one disputant has significant or highly significant observational evidence that implies the other's privileged self-knowledge claims is questionable, then don't trust that claim according to the following credence levels: highest signs = lowest credence; significant signs = low credence; little signs = high credence; no signs = highest credence.

Bergmann's and Christensen's views can't account for disagreements over privileged claims

Bergmann's, Christensen's, Sosa's, Lackey's, and Kelly's views can't adequately account for the examples of everyday peer disagreements about privileged self-knowledge described in the Introduction; they can't account for the examples from criminal law and from mental health therapy. The examples of everyday peer disagreements about privileged self-knowledge claims assume that one can scrutinize the mental states of others indirectly in ways discussed above. The five scholars of peer disagreement in their views of peer disagreement uphold implicitly the Inscrutability Thesis which says one can't scrutinize the mental states of others. As described in the last chapter, Christensen reasons that the person in the Careful Calculation case study can stay confident in her belief preferring her own evidence from privileged self-knowledge claims over the other's evidence from equivalent privileged self-knowledge claims just because a person can't scrutinize the other's comparable evidence to see how it stacks up to her own. As we have seen, the other four scholars have similar argument for steadfastness. None of them allows for the indirect scrutiny that is so common in everyday life.

People often make mistakes in their privileged self-knowledge claims even though those claims under normal conditions deserve prima facie true status. They say they know they are not biased against a different race when their actions show that they likely are; people often tell a significant other "I love you" when they actually are just infatuated; people often say they know they are not jealous when they really are. Of course, behaviors aren't always correlated with mental states. But they often are.

Christensen even says that the reliability assessment essential for determining who in a disagreement is more likely to get it right about the issue can be done even before the peer disagreement starts, as we saw in the last chapter. This strategy would obviate the value of the feedback we can get from others as regards to whether one's actions match one's claimed mental states, and in principle it assumes that one can't scrutinize the other's privileged self-knowledge claim.

There is another way in which Bergmann doesn't adequately recognize the fragile nature of privileged self-knowledge claims given the psychological, cultural, and physical barriers to the propositional truth of such claims. Bergmann uses the following privileged self-knowledge claim in order to justify extreme confidence that p where p is "Jack's behavior is morally wrong": "I am extremely confident in my insight about the way things really are which include p." He uses a paradigm example of how one can establish a strong justification of a belief in a similar way, namely, the belief that objects can be inferred from perception. He argues that just like we can establish a strong belief based on the insight about the way things really are as including objects can be inferred from perception, so too "Jack's behavior is morally wrong" can be strongly based on the insight about the way things really. The problem is that the former is much more secure than the latter in terms of justification. He doesn't see the comparatively fragile nature of the everyday privileged self-knowledge claims.

But even apart from this issue of the levels of secure propositions, another similar view of his also doesn't seem to account for the fragile nature of privileged self-knowledge claims. He believes that "extreme confidence" can add some justification to a claim, as we saw. Consider this: Based on extremely confident eyewitness testimonies, 375 people in the United States have been exonerated by DNA testing up to the year 2022, including 21 who served time on death row, reported by the Innocence Project. It is widely recognized that the claim to high confidence in a witness isn't highly correlated to accuracy (Roediger, 2012). People make mistakes about what they think they know about themselves even when they are extremely confident. Claims to high confidence about privileged claims are fragile even though they are usually correct.

Concluding remarks

The two norms were developed in order to detect, diagnose, and correct the misuse of privileged self-knowledge claims in peer disagreements. Such claims are misused when they are taken as invulnerable and inscrutable. Privileged self-knowledge claims are inherently fragile due to the many unconscious and conscious psychological, cultural, and physical barriers to self-knowledge, such as the unconscious self-enhancement motive (discussed in Part One), which encourages individuals to emphasize and prefer information gathered

and beliefs about oneself that make one feel good about themselves. Ulterior motives like self-enhancement don't care about the truth about oneself.

While the norms were developed in terms of Bergmann's misuse of privileged self-knowledge claims, they help us point out overreaches of privileged self-knowledge claims in every peer disagreement. The five scholars of peer disagreement we focused on think such claims can't be scrutinized. Bergmann says one disputant's privileged self-knowledge can't be defeated by another just because it can't be scrutinized. Yet they can be indirectly scrutinized. When a person claims privileged self-knowledge but their observable behavior and speech imply their claim is false, we appropriately think the claim is unwarranted. We appropriately normally trust the privileged self-knowledge claims of others, but they are fragile, often their propositional content is false. They are scrutable because they are fragile. Luckily, we can detect that fragility through indirect scrutiny.

PART THREE

Know Yourself Better in Peer Disagreements with the Assessment Framework

Chapter Six of Part Three accomplishes the objective that I started the dissertation with, namely, to derive a framework for assessing the comparative epistemic status of the claimant and the claim-critic in a peer disagreement about a privileged self-knowledge claim. The framework must incorporate an understanding of the fragile nature of privileged self-knowledge claims, and it must provide a way of recognizing when a privileged self-knowledge claim may not reflect the actual mental state of the claimant. It must include an understanding of how the observational-interpretive and privileged access methods of acquiring self-knowledge can work together with mindfulness to make privileged self-knowledge claims more reliable and to show each disputant what their own likelihood is of getting the issue right. It must remedy the state of the scholarly literature on peer disagreement which we found doesn't adequately take into account the fragile nature of privileged self-knowledge claims in the comparative assessment of each disputants' epistemic standing. It must be able to account well for the complexity of peer disagreements about privileged self-knowledge claims.

The Assessment Framework derived in the next chapter does all these things. All the six key components of peer disagreements about privileged self-knowledge claims are incorporated into this Assessment Framework along with the four interrelations discovered. The Assessment Framework makes it possible to know yourself better in peer Disagreements.

Chapter Six

Better Assessing Privileged Self-knowledge Claims in Peer Disagreements With the Assessment Framework

This chapter argues for the following research-based Assessment Framework for evaluating privileged self-knowledge claims when they are crucially consequential in peer disagreements:

> **Assessment Framework:**
> In a peer disagreement where the privileged self-knowledge claim of one disputant is crucially consequential for the disagreement, trust the claim prima facie only if there are no or little significant signs of "judgmental awareness" and/or of observational evidence that implies the claim is questionable; and adjust credence in the privileged self-knowledge claim according to the following scale of such significance: No signs = highest credence; little signs = high credence; significant signs = low credence; highest signs = lowest credence.

We derive and justify this Assessment Framework from two sources: from the following six key components found in previous chapters to be central components of target disagreements, and from the four following interrelations of those components:

> **Key components in peer disagreements about privileged self-knowledge claims:**
> 1) The fragility of privileged and observational-interpretive access due to the many psychological, cultural, and physical barriers.
> 2) The observational-interpretive access method for gaining self-knowledge of mental states in peer disagreement.
> 3) The privileged access method for gaining self-knowledge of mental states in peer disagreement.
> 4) Mindfulness
> 5) The Prima Facie Norm.
> 6) The Indirect Scrutability Norm.
>
> **Four interrelations among the six key components**
> *Interrelation #1: The two norms are interrelated because the two methods of knowing oneself are interrelated.*
> *Interrelation #2: Mindfulness improves both the methods of knowing oneself.*
> *Interrelation #3: Each disputant can observe signs of higher and lower mindfulness.*
> *Interrelation #4: Each disputant can have four different types of observational-interpretive evidence.*

These components are central components of target disagreements. The components and their interrelations imply the Assessment Framework.

The scare quotes around "judgmental awareness" in the Assessment Framework description require explanation. "Judgement" here doesn't refer to specific judgments that we have to make in our daily lives, e.g., that the checker is competent, that I will be safe walking down this street at night, etc. Rather, "judgmental awareness" refers to the particular quality of the awareness in which everyday judgments are formed, that is, an awareness in which one is rash to judge what one is aware of.

We can understand "judgmental awareness" better by remembering what we have seen in Part One about the nonjudgmental awareness of mindfulness. In Part One we found that the core description of mindfulness is nonjudgmental awareness. Judgmental and nonjudgmental awareness measure the same thing. Signs of judgmental awareness are signs of lower mindfulness, and signs of nonjudgmental awareness are signs of higher mindfulness. And we discovered that judgments and self-knowledge claims made in the context of nonjudgmental awareness more reliably reflect the reality of one's mental states. This is so because the nonjudgmental awareness of mindfulness facilitates many self-knowledge-conducive benefits, more balanced information about the self since negative information about the self is less of a threat; less barriers to self-understanding since the ulterior motives (e.g., self-enhancement motive) aren't needed; more flexibility in one's thinking since perspectives and interpretations are seen as interpretations rather than solidly fixed realities. In the non-judgmental awareness of mindfulness, one is able to reassess longstanding and automated judgments or self-knowledge claims because mindfulness's act of seeing them as interpretations creates the interruption and consequent reflective distance that allows them to be reassessed. Thus, with nonjudgmental awareness, mindfulness, one can engage deliberations about oneself free of the influences of automated judgements and equipped with balanced information about oneself.

We can now see why it makes sense in the Assessment Framework to say: Trust the claim prima facie, only if there are no or little significant signs of "judgmental awareness." The following are signs of judgmental awareness and its consequent effects on self-knowledge claims:

Signs of judgmental awareness … and its consequent effects on self-knowledge claims include:

When the claimant:

1) Makes automated judgements about what one is aware of,
   - Leading to new information not being factored in.
2) Is strongly reactive to negative information without adequately considering it,
   - Pointing to a defensive mechanism motivated by self-preservation rather than by the desire to know the self.
   - Indicating that negative information is threatening.
   - Indicating ulterior motives are needed.
3) Doesn't see her self-knowledge claim as an interpretation,
   - Lacking the reflective distance to adopt more appropriate and healthy understandings of oneself.
4) Can't even entertain negative information about oneself,
   - Indicating the claim is made without a realistic balance of information.
   - Leading to a grandiose assessment of one's skills.
5) Judges others/oneself/things without adequately considering information that goes against the judgment,

- Indicating judgements made rashly.
- Lacking diverse options for judgments.

Such signs of judgmental awareness can be appropriately used as evidence for not trusting the privileged self-knowledge claim prima facie. Such signs of automated judgments of things in awareness are evidence that the claimant has asserted her privileged self-knowledge claim in a context not conducive to such claims reliably reflecting the reality of her mental states. It is for these reasons that the Assessment Framework appropriately recommends not trusting the claimant's privilege claim about herself prima facie when these signs exist. We will see how one can actually measure whether one is predisposed to these particular signs of judgmental awareness with the Five Facet Mindfulness Questionnaire, and this can give us an indication of the likelihood of the person to get their privileged self-knowledge clam right. Thus, it makes sense in the Assessment Framework to have the trust of the prima facie status of a privileged claim depend on whether there are signs of judgmental awareness.

I start first by pointing out four crucial interrelations derived from these six key components. At the same time, I show how the four interrelations and the six components are embedded in the Assessment Framework. Next, I test the Assessment Framework to see how well it works to explain what is going on phenomenologically in a diverse variety of examples of typical peer disagreements about self-knowledge claims.

Four interrelations of the six components and how they are embodied in the Assessment Framework

*Interrelation #1: The two norms are **interrelated** because the two methods of knowing oneself are interrelated.*

Let's first remind ourselves what these two methods are referred to in the statement of the first interrelation stated in the heading just above, and then we will see how they are interrelated. The two methods of knowing oneself are the method from observational-interpretive access and the method from privileged access. In Part One I showed how empirical studies on self-observation and confabulation prove that we <u>often</u> do know our own mental states through sensory and interpretive access. For example, people come to think that they are happy when they find themselves smiling, they are hungry when they hear their stomach growling, or they are attracted to a person when they observe themselves blushing or nervous.

And this observational-interpretive way of knowing oneself is used to know the mental states of others. I come to think my friend is happy, when I see her smiling; hungry, when I hear her stomach growling; attracted, when I observe her blushing and nervous. We need this quick observational-interpretive method of knowing our own mental states, and those of others, in order to function properly in our lives, as we have discussed in Part One. Beliefs about oneself and others formed through observational-interpretive and interpretive access are reliably formed to the extent that observable cues trigger an interpretation of the mental state that reflects what the mental state really is.

Also, in Part One we saw how the method of knowing oneself through privileged access is very different than the one through observational-interpretive access. While observational-interpretive access typically is unconscious, interpreted, and sensory based, the method of knowing oneself through privileged access is conscious and deliberative. The decision or judgment resulting from privileged access settles an issue through considering alternative possible answers to the issue and then making up one's mind by engaging one of the possibilities. The propositional attitudes that result from the deliberative agency are known immediately and directly without need for observation.

We have found any claim to self-knowledge is fragile. Observational-interpretive access fails to yield self-knowledge when cues are ambiguous and when ulterior unconscious motives encourage interpretations of cues that don't reflect the mental states people actually have, as discussed in Part One. Even the most careful deliberation and decision about one's intent of an action can be mistaken when unconscious defense mechanisms are deployed to protect one from the dysphoria that would result were one to know one's actual intent. In Chapter Four we pointed out mistakes some of the best scholars of peer disagreement make about privileged self-knowledge claims, mistakes they make both because they are not adequately aware of the fragility of self-knowledge claims and because they are not aware of the integration of the two methods. We have consequently come to know that all privileged self-knowledge claims are fragile due to all the possible psychological, physical, and social barriers to self-knowledge.

It is exactly because of this fragility that the two norms we have recognized are needed. The Prima Facie Norm recommends granting privileged access prima facie, but only to the extent that there are no good reasons

to discount it due to its fragility. The Indirect Scrutability Norm is designed to point out when a privileged self-knowledge claim should not be trusted due to this fragility.

We can now say why the two norms we have uncovered are interrelated. They are interrelated because the two ways of knowing are interrelated. The interrelation of the two methods of knowing oneself is based on the fact that one's mentals states typically influence one's behaviors in a way consistent with the mental states. What this means is that I and others can reliably know indirectly what my mental state is from my behavior, even though only I can know what my mental state is directly through my privileged access. While I can't know in a first-person way what another person's mental state is, I can know what that person's mental state is indirectly through the observation of that person's behavior, and vice versa.

The interrelation of the two ways of knowing here is the basis for the interrelation of the Prima Facie Norm and the Indirect Scrutability Norm. The Indirect Scrutability Norm says that one can have evidence from a person's behavior that implies the person's privileged self-knowledge claim is false. The Indirect Scrutability Norm is related to the Prima Facie Norm in that the former through observation-based can scrutinize indirectly the unobservable privileged self-knowledge claim. This relation between the two norms is only possible because of the following relation between the two ways of knowing oneself: The observational-interpretive method of knowing oneself can uncover observed evidence that indirectly scrutinizes the unobservable results of the privileged method of knowing oneself. This interrelation between the two ways of knowing oneself is possible only because a person's mental states typically influence the person's behavior such that if the behavior is inconsistent with the mental state, this is evidence the person doesn't have the mental state. Hence the claim-critic can only indirectly argue that the claimant doesn't have the mental state she claims. And this is so because the claim-critic only has evidence from the claimant's behavior and speech that is incompatible with the mental state claimed by the claimant, and because the claim-critic uses this observational-interpretive evidence to argue that the claimant doesn't have the mental state she claims. It is an indirect argument which presupposes the indirect scrutiny of the mental state claimed by the claimant, which mental state the claim-critic can't directly observe.

The interrelation described above between the two norms is embodied in the Assessment Framework. The key formula of the Assessment Framework says, "Trust the claim prima facie, only if there are no or little signs of 'judgmental awareness' and no or little significant signs of observational-interpretive evidence of behaviors and speech that implies the claim is questionable." This key formula for assessing how far to trust particular privileged self-knowledge claims expresses a necessary condition for granting prima facie status, namely, that there is no observational-interpretive evidence that implies the claim is unjustified. The claimant says yes there is no observational-interpretive evidence that implies the claim is unjustified, and so one should grant prima facie status. And the claim-critic says no, the necessary condition for prima facie status isn't present.

This key formula of the Assessment Framework is a combination of the two norms. And it can be a combination of the two norms because of the interrelation pointed out above, namely, that the Indirect Scrutability Norm can indirectly scrutinize through observational-interpretive evidence the unobservable mental state pointed out by a privileged self-knowledge claim referenced by the Prima Facie Norm. I can clarify why it is said in the last sentence that the Indirect Scrutability Norm can indirectly scrutinize through observational evidence the unobservable mental state. For the person who doesn't have the mental state claimed, the claim-critic, that person can only have indirect observation of the claimant's mental state based on an inference about the claimant's mental state; and this is so because only the claimant can have privileged direct access to her mental state. This inference is based on observations of the speech and behavior of the claimant relevant to her claimed mental state. If these observations are compatible with the claimed mental state, it is inferred that the person has the mental state, and if these observations are incompatible with the claimed mentals state, it is inferred that the claimant doesn't have the mental state. So, the claim-critic in this sense has <u>indirect</u> observational-interpretive access to the unobservable mental state of the claimant to the extent that she has observational-interpretive evidence of the claimant's behavior relevant to the mental state in question. Like Carruthers the observational-interpretive access to one's mental state and the mental states of others comes not from observing the mental state itself, but rather from observing the behaviors, speech, countenance, etc. of others and oneself. My view entails that the claimant has both indirect observational-interpretive access and direct privileged access to her mental state.

We can see now how the key formula of the Assessment Framework combines the two norms in order to put to use the interrelatedness of the two norms. The Prima Facie Norm says, "Trust a disputant's privileged self-knowledge claim prima facie if there are no or little signs that it is unjustified." The signs of a claim being unjustified are exactly the situation pointed out in the Indirect Scrutability Norm, which says, "If one disputant has observational-interpretive evidence that implies the other's privileged belief about the self is unjustified, and shares the observational-interpretive evidence with the other, then the other's justification for this privileged belief is defeated." The good reason to think the prima facie status should not be granted to the privileged self-knowledge claim is that there is observational-interpretive evidence that implies this claim is unjustified. It should now be clear how the key formula of the Assessment Framework embodies the interrelation between the two norms.

Interrelation #2: Mindfulness improves both methods of knowing oneself.

I already discussed thoroughly in Part One why and how increases in mindfulness make privileged self-knowledge claims more reliable. A tremendous amount of empirical research proves this point, namely, that higher mindfulness is positively correlated with higher reliability of privileged self-knowledge claims.

What we haven't said yet is that the claimant can appeal to observable signs of mindfulness to support her privileged self-knowledge claim. While I will give an example of this in the next section, I can say what this situation would look like. Suppose a person has a privileged self-knowledge claim. Another person contests this claim based on observational-interpretive evidence. At this point the claimant can describe her own observations of herself asserting that her behaviors are in line with the mental state referred to in her claim. And the claimant can appeal to observable signs of mindfulness to support the reliability of her claim. She points to behavioral patterns that are positively correlated with mindfulness, like enhanced observation and a non-judgmental awareness. If one's behavioral patterns indicate one is higher in mindfulness, and higher mindfulness indicates one's privileged self-knowledge claims are more reliable for reflecting the reality of one's mental states, then this observation of signs of mindfulness does lend more credibility to her privileged self-knowledge claim.

Undoubtedly, mindfulness improves observational-interpretive access, which improves both one's own self-knowledge and one's knowledge about another's mental state. It does this with its enhanced observation. Such access relies, whether geared towards self-knowledge or knowledge of the other's mental state, on information that is attained through perception, self-perception, proprioception, and interoception. The better the information is, the better one is able to understand what one's mental states are and what the mental states of others are through observational-interpretive access. The better the information the better one is at recognizing observable cues embodying interpretations which, when triggered, activate an understanding of one's mentals states or one's understanding of the other's mental state.

Mindfulness also improves self-knowledge and knowledge of others by reducing biases in one's observational-interpretive access to one's mental states and the mental states of others. For example, as seen in Part One, mindfulness has been proven to reduce the dysphoria that results from recognizing one's biases, whether they are biases to see oneself as better than the other or to see the other as worse. Also, mindfulness helps one to avoid overestimating one's own skill; Dunning-Kruger research has proven people have a tendency to overestimate one's own skills (Dunning, 2003, 2004, 2005). Mindfulness does these things through its non-judgmental way of observing. The more one can tolerate due to non-judgmental observation the dysphoria produced by being open to information that indicates one's skills aren't as good as one thinks or information that indicates one is biased, the more likely one is to know one's actual mental states. Since mindfulness helps one have a more realistic understanding of oneself through observational-interpretive access, and since a more realistic understanding of oneself often helps one know the mental states of others through observational-interpretive access by comparing the other to a more realistic understanding of oneself, mindfulness helps one have more realistic beliefs about other peoples' mental states. To see this, think about the ways in which you would have a better understanding of a person's intentions and motives if you come to find out as a result of mindfulness that you likely have an inappropriate bias against that person, or if you come to find out due to the more realistic information facilitated by mindfulness that you likely have an overinflated understanding of your own skill level making you predisposed to think inappropriately that your colleague's skill level is lower than

yours. Thus, mindfulness increases the reliability of the observational-interpretive access method for determining one's mental states and the mental states of others.

The Assessment Framework embodies the discovery that mindfulness improves both privileged and observational-interpretive access. And it does this for both the claimant's attempt to get the privileged self-knowledge claim right and for the claim-critic's attempt to get right the criticism of the claim based on observational-interpretive access. The Assessment Framework recognizes the factor of mindfulness so much so that signs of inadequate mindfulness are evidence for not granting prima facie truth status to privileged self-knowledge claims. The idea here is that the lower the mindfulness level, the less likely one is to claim privileged self-knowledge that actually reflects the reality of one's mental state. Also, in making the absence of inadequate levels of mindfulness a necessary condition for granting prima facie status, the Assessment Framework opens the door to using the level of mindfulness observed of the claim-critic as a criterion for how seriously to consider the claim-critic's observational-interpretive evidence against the claimant's privileged claim. Similarly, the thinking here is that the higher the levels of mindfulness observed in the claim-critic's behavior and speech, the more likely the claim-critic's observations are free of bias or self-enhancement ulterior motives.

Interrelation #3: Each disputant can observe signs of higher and lower mindfulness.

Having an Assessment Framework that assigns credence levels of claimant views versus claim-critic views depending significantly on the level of mindfulness they have is useless if there is no clear way to decide what a person's mindfulness levels are in peer disagreements about privileged self-knowledge claims. Luckily, there are two ways to gauge someone's mindfulness levels, through observation of personality traits associated with higher or lower mindfulness and through observation of behaviors known to be closely associated with particular mindfulness levels. And, as we have said, with an estimate of someone's mindfulness level, one can infer the person is more or less likely to have accurate claims about her mental state, whether that claim is based on privileged access or observational-interpretive access.

Empirical studies on mindfulness show that certain personality traits are indicative of higher levels of mindfulness, while others are indicative of lower levels. For example, recent research shows that there is a positive correlation between mindfulness and the personality traits of adjustment and ambition. So, seeing these

traits in someone is some evidence that the person has higher mindfulness levels. On the other side the personality traits of caution, leisure, and excitability are negatively correlated with mindfulness (Altizer, Ferrell, and Natale, 2021) (Travers, 2020), (Hanley, et alia, 2015) (Hanley and Garland, 2017). Narcissism as well has been demonstrated to be negatively correlated with mindfulness. To the extent that one can estimate reliably these personality traits in a person, to that extent one can get some evidence for how likely the claimant or the claim-critic is to get the issue right, since it has been proven that mindfulness improves observational-interpretive and privileged access.

There is another way of estimating claimant and claim-critic mindfulness levels, through signs of behavior tendencies known to be closely associated with particular mindfulness levels. Researchers on mindfulness have produced mindfulness questionnaires asking subjects in their studies to self-report answering questions about whether they have behavioral and thinking tendencies closely associated with particular mindfulness levels. These mindfulness measurement scales ask subjects to observe and remember their behavior and thinking in order to determine if they have the specific behavioral and thinking tendencies closely associated with particular mindfulness levels. The mindfulness scale we will be using, the Five Facet Mindfulness Questionnaire (FFMQ), has been demonstrated to have validity and reliability for accurately pointing out particular mindfulness levels based on particular behavior and thinking tendencies. To the extent that claimants and claim-critics in peer disagreements about privileged claims are aware of these observable behaviors and thinking tendencies, they can use them to gauge the levels of mindfulness and so the accuracy of the evidence presented in the peer disagreement. For example, when the claimant's privileged self-knowledge claim is in dispute, she could present to the claim-critic evidence of higher mindfulness given that she has many of the tendencies associated with higher mindfulness, and this in turn can be some evidence that her privileged self-knowledge claim reflects the reality of her mental state. And the claim-critic could present to the claimant evidence of higher mindfulness given that he has many of the tendencies associated with higher mindfulness, and this in turn can be some evidence that his critique of the claimant's claim is based on accurate observations. On the flip side, claimants in such disagreements can find evidence of the claim-critic's low mindfulness and have some evidence that his critique isn't based on accurate observations. Of course, the mindfulness

measurement scales are meant to be used by subjects of empirical studies for self-reporting based on their observations of themselves. But, once one knows what behavioral and thinking tendencies are associated with particular levels of mindfulness—and this isn't hard to do—one can in such disagreements look for those tendencies in their opponents and subsequently have some evidence for how to weigh the other's views in comparison to one's own.

To see how this would work, we can take a look at the most respected and most used mindfulness measurement scale, the Five Facet Mindfulness Questionnaire (FFMQ) developed by Ruth Baer and colleagues. The questionnaire measures five different aspects of mindfulness: observing, describing, acting with awareness, nonjudging of experience, and nonreactivity. The observing facet measures the individual's tendency to be aware of details in one's environment, proprioception, interoception, thoughts, feelings, etc. The describing facet measures the subject's capacity to name, categorize, and describe thoughts, feelings, etc. Acting with awareness measures the individual's ability to stay present and aware while acting in the world. The nonjudging of experience facet measures the ability of the person to have nonjudgmental considerations of thoughts and feelings. And it measures the extent to which people tend to have automated judgments for particular experiences. The nonreactivity facet measures the subject's propensity for remaining calm, not ruminating, and not reacting automatically to thoughts and situations that typically would elicit negative responses.

Furthermore, the claimant and the claim-critic can use a measure of mindfulness like the FFMQ to help look for signs of judgmental awareness. The FFMQ for example has specific questions that the claimant and the claim-critic can use for identifying judgmental awareness. These questions are proven to be negatively correlated to mindfulness. There are seven questions used to identify judgmental awareness in the form of reactivity, seven questions to gauge the level of acting with awareness, seven questions for identifying the awareness of a person, and seven to determine the level of judgmental experience. In using these questions claimants and claim-critics have a way of gauging the judgmental awareness the person has, consequently the mindfulness level a person, and in turn the likelihood that a person gets the issue correct. We will see how these specific questions can be used in the case studies below.

The Assessment Framework developed here incorporates a provision for measurements of mindfulness that can make a difference in the assessment of whether to grant prima facie true status to privileged self-knowledge claims in peer disagreements. Of course, many people in peer disagreements about privileged self-knowledge claims aren't aware of what mindfulness is. Even so, people have always gauged how seriously to weigh the perspectives of disputants in peer disagreements about privileged self-knowledge claims in terms of behavioral and thought tendencies such as the ones that are central to mindfulness even without knowing that these tendencies are central to mindfulness, indeed without even knowing what mindfulness is. For example, you would expect someone to take less seriously the view of a disputant about a privileged self-knowledge claim if that disputant clearly through observation is shown to rashly jump to conclusions with knee-jerk emotional reactions, or the disputant shows herself to automatically judge a situation without considering the nuances of the situation. And people in peer disagreements about privileged self-knowledge claims may not know the other person enough to be able to say what level of mindfulness the person is at. It is just that people often are in a position to know whether the other is highly reactive, highly judgmental in an overly automated way, or not sufficiently aware of fine grain details that can be decisive for resolving a disagreement, especially if this is a peer disagreement. And when they are, and they realize these tendencies are negatively correlated with the accuracy of one's claim to self-knowledge, mindfulness (whether they know this name or not) can help more accurately assess who in the disagreement is more likely to have the right view. Knowing more about the details of mindfulness studies, which specific behavioral tendencies are indications of higher or lower mindfulness, which questions to ask to determine mindfulness levels, how strongly behavioral and thinking tendencies are correlated with the accuracy of privileged self-knowledge claims (the r value tells one how strong the correlation is with a value from -1 to 1) can help one determine more precisely whose view, the claimant's or the claim-critic's, in the peer disagreement about privileged self-knowledge claims is more likely to be right. We will see how this can help in the case studies in the next section.

Interrelation #4: Each disputant can have four different types of observational-interpretive evidence

It is extremely important to understanding that both the claimant and the claim-critic have four different types of observational-interpretive evidence. We have referred to all four types of observational-interpretive

evidence in this chapter. Here we explicitly list them below. It is with these four different types of observational-interpretive evidence that the claimant and claim-critic will come to their assessment about who is more likely to get the issue right. Both the claimant and the claim-critic have observational-interpretive evidence of their own behaviors, thoughts, and emotions. Both can have observational-interpretive evidence of the other's behaviors, thoughts, and emotions. Both can have signs of their own judgmental awareness. And both have, or could have, signs of the judgmental awareness of the other. All these four types of evidence each can come into play for deciding who is more likely to get the issue correct. We will see in the next section how all four types of observational-interpretive evidence are used in the case studies below. The following is a pithy summary of the four different types of observational-interpretive evidence each disputant has for a total overall between the two disputants of eight types:

Eight types of observational-interpretive evidence for the assessment of the privileged self-knowledge claim:

Claimant perspective

Observational-Interpretive evidence the claimant has of

1) The claimant's own behaviors and speech especially whether this evidence "implies" a false claim.
2) The claimant's own signs of mindfulness especially as this relates to the reliability of the claim.
3) The claim-critic's behaviors and speech especially as related to the evidence she brings.
4) The claim-critic's signs of mindfulness especially as related to the reliability of the evidence she brings.

Claim-critic perspective

Observational-Interpretive evidence the claim-critic has of

1) The claim-critic's behaviors and speech especially as related to the evidence she brings.
2) The claim-critic's signs of mindfulness especially as related to the reliability of the evidence she brings.
3) The claimant's behavior and speech especially whether this evidence "implies" a false claim.
4) The claimant's signs of mindfulness especially as this relates to the reliability of the claim.

Of course, the claimant has the privileged self-knowledge claim, and the claim-critic doesn't. The claimant could just forget about any of the evidence the claim-critic has. But, if the claim-critic is truly an epistemic peer, the claim-critic isn't going to just present frivolous observational-interpretive evidence that implies the claimant's privileged self-knowledge claim doesn't reflect the reality of the claimant's mental state. No, if that person is truly an epistemic peer, the observational-interpretive evidence presented is very challenging. Consequently, the claimant can't just claim over and over the truth of her privileged self-knowledge claim. No, she has to at least take seriously the evidence the claim-critic brings.

In such disagreements we are concerned with the signs of inadequate mindfulness, namely, judgmental awareness. The judgmental awareness referred to in the Assessment Framework tends to function as a way of gauging how seriously to weigh the observational-interpretive evidence. In such disagreements a claim-critic

can present evidence that would definitively imply the privileged self-knowledge claim is false if the evidence really does exist. The higher the mindfulness, the more reliable the privileged self-knowledge claim actually reflects the mental states the claimant actually has. Each type of observational-interpretive evidence will factor in to each disputants' understanding of who is more likely to get the issue right. And they will debate the significance of each type of observational-interpretive evidence.

The Assessment Framework does a very good job of taking account of, and incorporating all of, the four different interrelations and the six components the interrelations are based on. It provides a shared framework for each disputant to rely on so that they can come to an assessment of the privileged self-knowledge claim in dispute. It provides a framework for disputants to discuss and debate the observational-interpretive evidence that is presented either from the claimant or from the claim-critic. The Assessment Framework tells one where the observational-interpretive evidence is crucial, and how it is crucial, and how it factors into the assessment. The Assessment Framework tells us how each disputant's respective four types of observational-interpretive evidence fit into the equation that determines whose view is more likely to be right.

Evaluating the Assessment Framework

We can see now how well the Assessment Framework accounts for the phenomenology of four typical examples of peer disagreement over privileged self-knowledge claims. And I have a standard by which to measure the success of this assessment tool:

> **Assessment Framework:**
> In a peer disagreement where the privileged self-knowledge claim of one disputant is crucially consequential for the disagreement, trust the claim prima facie only if there are no or little significant signs of "judgmental awareness" and/or of observational evidence that implies the claim is questionable; and adjust credence in the privileged self-knowledge claim according to the following scale of such significance: No signs = highest credence; little signs = high credence; significant signs = low credence; highest signs = lowest credence.

In what follows we will work towards judging how well the Assessment Framework meets this standard from both the perspective of the claimant and the claim-critic. In doing so we will present two case studies that reflect the claimant's perspective, one that strongly and intuitively favors a conciliatory credence and another that strongly and intuitively favors a steadfast credence. And next we will present two case studies that reflect the claim-critic's perspective, one that strongly and intuitively favors a conciliatory credence and another that

strongly and intuitively favors a steadfast credence. After each case study we will evaluate whether the Assessment Framework organically helps us make sense of the intuitions.

Know Yourself in disagreements from the claimant's perspective

To see how well the Assessment Framework accounts for typical case studies, we will be using the four case studies presented in the introduction to this dissertation. We will consider peer disagreements about privileged self-knowledge claims from the only two perspectives that matter, the perspective of the claimant and the claim-critic. We will first consider how claimants know themselves in peer disagreements over their own privileged self-knowledge claims.

When the strong intuition is that the claimant should conciliate

We first see to what extent the Assessment Framework accounts for, and helps us understand better, a case study where there is a strong intuition that the claimant should conciliate. Let's see how well the Assessment Framework explains how the JEALOUSY case study works, and here is the case study again:

> **JEALOUS: "I know I am not jealous!"**
> Kamal talks to his longtime therapist about difficulties he has been having with a person at work whom he hates because this person is mean. He sees his therapist as an impartial epistemic peer on interpersonal issues. He describes the many ways that this person is mean in different cases. The therapist tells him that she thinks he, rather, hates the person because he's jealous. Kamal's irritation is clear when he says, "I know I am not jealous!" His therapist says,
>
> > "You describe that your colleague has achievements that you value and have not accomplished. We have talked about some instances in the past when you conceded jealousy when initially you insisted you weren't. Sometimes you don't pay attention to how your emotions influence your thoughts. You say you are not a detail person; you don't like to pay attention to distractions like clocks ticking, birds chirping, and smells of things. You say when driving you don't care to notice things if they aren't matters of safety. I don't think you would be able to observe how your emotions influence your thoughts in this stressful situation. I can see them because I am, unlike you, impartial in this case. Furthermore, your reactivity to the idea that you are jealous was impulsive without hearing me out or even taking the view seriously; that may indicate you don't want to hear the evidence, that it is a defense mechanism triggered to help you think good things about yourself."
>
> Kamal remembers the incidents she is talking about where he finally conceded jealously, but he thinks this case is significantly different. Sometimes his therapist is wrong. While he thinks the past instances of similar mistakes are inconclusive, he finds compelling the point that he often doesn't see the extent to which his emotions influence his view of others, given that he isn't a detail person. And it is telling how he was so reactive. He has less confidence in his view.

Notice here how the observational-interpretive evidence of the therapist about his past mistaken understandings of his motives is challenging for Kamal, since it does imply that the privileged self-knowledge claims of Kamal sometimes don't reflect reality. But Kamal does eventually defuse the power of these examples from the past that imply he likely doesn't have the motive he thinks he has, given that he has examples for when he thinks his therapist was wrong. His observational-interpretive evidence of his own behaviors counteracts the observational-interpretive evidence the claim-critic gathers from his behaviors.

What does end up causing Kamal to reduce his credence level is the two signs of inadequate mindfulness that his therapist cites. The therapist points out the significant reactivity of Kamal and his historical inability to see how his emotions influence his thoughts about a situation and others. In fact, the FFMQ asks people questions about 1) whether they recognize how emotions influence one's thoughts about others, and about 2) whether they have knee-jerk and strong reactions to interpersonal issues; these two types of questions on the FFMQ are associated with two facets of mindfulness, nonjudgment and reactivity respectively. When the subject answers no to the first question and yes to the second, as Kamal does, the FFMQ considers these signs of lower levels of mindfulness. Recognizing that his difficulty seeing how his emotions influence his assessment of others impacts negatively his ability to correctly judge what his motive is for the hatred for the man at the office, and recognizing also that his reactivity may well be a sign of a defense mechanism hiding the truth of his motive for hatred, he then thinks his privileged self-knowledge claim may not reflect his actual motive. Kamal doesn't know what mindfulness is, but that doesn't mean that he can't apply the Assessment Framework. Unbeknownst to him, he intuitively looks for signs of mindfulness deficiency—that is, judgmental awareness—when he is trying to observationally verify his privileged self-knowledge claim. And it makes sense that he is looking for the same signs of mindfulness deficiency as is the FFMQ, since many of the characteristics of mindfulness harmonize well with the knowledge of cognitive impairment that psychologists have known long before mindfulness became popular in the West. It is commonly understood, for example, that emotions influence thought about one's motives, and that this can mean that one is less likely to accurately understand his motive for hatred. But what he inadvertently is doing is, first, finding that he doesn't have adequate levels of mindfulness, and, second, recognizing that this deficiency provides some evidence that his privileged claim is unjustified. The necessary condition for granting prima facie truth to a privileged self-knowledge claim isn't met, and so he no longer can take for granted that his privileged claim is true.

This organic way in which the Assessment Framework handles this case is testimony to the usefulness of this Assessment Framework for peer disagreements about privileged self-knowledge claims. No other known assessment technique given in the peer disagreement literature accounts for "signs of inadequate mindfulness" as a measure influencing credence levels. This case study strongly and intuitively favors a lower credence level

for Kamal, and that is what the Assessment Framework organically delivers. Furthermore, the Assessment Framework explains exactly why a more conciliatory stance makes sense, because a necessary condition for trusting the privileged self-knowledge claim isn't met, and it isn't met because the mindfulness level is inadequate.

One other important benefit of this Assessment Framework must be pointed out. According to the Assessment Framework, the necessary condition needed for trusting the privileged self-knowledge claim is the following: There are no signs of judgmental awareness and no observational-interpretive evidence of behaviors and speech known to each that implies the claim is unjustified. In this case study the necessary condition is missing only because there are signs of inadequate mindfulness. There isn't observational-interpretive evidence of behavior and speech that implies the claim is unjustified because the therapist's observational-interpretive evidence that Kamal makes mistakes about his motives is canceled out by Kamal's observational-interpretive evidence of his therapist making mistakes. This means that the Assessment Framework developed here has the ability to differentiate between different ways that the necessary condition for trusting prima facie is missing.

Notice that the therapist and Kamal use all of their four different types of observational-interpretive evidence listed above in their evaluation of Kamal's privileged self-knowledge claim. Both remind themselves of the times they observed Kamal's speech and behavior concluding that they imply he sometimes doesn't see how his emotions influence how he thinks about others. Both remind themselves of the times they observed the therapist's speech and behavior concluding that they imply he sometimes misunderstands Kamal's mental states. Both consider signs of judgmental awareness by observing Kamal's reactions. Both also look for signs of judgmental awareness by observing the therapist's reactions. For example, both observe that the therapist is impartial.

When the strong intuition is that the claimant should be steadfast

We can continue to look at peer disagreements from the perspective of the claimant, and this time we present a case study that intuitively and strongly favors the steadfast response of the claimant. Again, the overall goal is to see if the Assessment Framework naturally accounts for complex peer disagreements over one disputant's privileged self-knowledge claim. Consider the following Love case:

LOVE: "I know I love you!"
Blake says to his wife Anna, "I love you." Anna says, "No you don't; many of the things that you do and say indicate that you don't love me; though I know that you sincerely think you do love me!" At this point Blake sees his wife as an epistemic peer. He is eager to hear her out because he knows her to be very sincere and perceptive. Anna describes how he yells at her often about her not being able to get along with her supervisors at work, degrades her for not being able to drive and insisting that he constantly taxi her and their son wherever she wants to go, complains about her obsessions, and berates her in public. While Blake acknowledges that he has done these things and that these things often indicate the absence of love, he believes there are extenuating circumstances. He says the acrimonious behaviors result from his extreme difficulty coping with the negative consequences—her not being able to keep a job to support a child, not being able to drive, obsessions, and excessive worries—that result from her three disabilities officially diagnosed by a psychologist: Attention Deficit Disorder of the Inattentive Type, Generalized Anxiety, and Delusional Disorder of the Non-Bizarre Type. He also thinks she isn't doing enough to ameliorate her difficulties, and to help others who have difficulties with her three disabilities. To his therapist he describes his acrimonious behavior, and he also describes how she always quickly reacts to his acrimonious behaviors in an automated and angry way without taking his concern for extenuating circumstances seriously, claiming he isn't a good person, just selfish. His therapist explains how such a knee-jerk reaction often is a sign of a defense mechanism which makes it less likely to correctly understand another's mental state by preventing consideration of alternative views or information which may be true and may soften one's criticism. At the beginning he reduced his credence level because she has such extensively good powers of reasoning. But in the end, he comes to be steadfast in reasoning both that she just isn't adequately able to see his legitimate extenuating circumstances, and that her interpretation of him as not a good person and selfish is influenced by her documented Delusional Disorder.

Notice here that Blake is ultimately steadfast about his privileged self-knowledge claim not primarily because he thinks his alternative explanation of his acrimonious behavior as due to extenuating circumstances is better than his wife's view. Rather, what motivates his steadfastness is two types of observational-interpretive evidence that he has: He has observational-interpretive evidence that Anna in effect has low levels of mindfulness demonstrated by high reactivity making her unable to give his alternative explanation its due consideration, and he has observational-interpretive evidence that her verbally expressed view of him as not a good person and selfish is influenced by her documented Delusional Disorder. Blake doesn't know that immediate reactivity and automated judgment are signs of lower mindfulness, but they are such signs nonetheless. Mindfulness isn't just a meditation practice developed by Buddhist monks 2500 years ago in order to verify that there is nothing permanent. Rather, the higher the level of mindfulness the more likely it is that the person gets the privileged self-knowledge claim correct. It is no accident that the cognitive processes people like Kamal and Blake come to identify with a little help from psychological research as inhibiting correct understanding of one's mental states are highly correlated with facets of mindfulness—like acting with awareness, nonjudging of experience, and nonreactivity. Buddhists have been experimenting for 2500 years with techniques and skills for understanding one's mental states better.

Those two very different types of observational-interpretive evidence Blake has about Anna's signs of inadequate mindfulness and observational-interpretive evidence of Anna's speech asserting he is selfish and not a good person block Anna's two efforts to make Blake unable to trust his privileged self-knowledge claim that he loves her. Let's describe the details of how this works. Anna first claims her own observational evidence

of Blake's behavior implies he really doesn't lover her. The observational-interpretive evidence she is referring to follows: Blake yells at her, degrades her, and berates her in public. This observational-interpretive evidence implies Blake's privileged self-knowledge claim to love her doesn't reflect the reality of his mental state. If this observational-interpretive evidence really does imply Blake doesn't love her, then Blake no longer has a necessary condition for trusting prima facie the truth status of his privileged claim to love her. And this would mean the necessary condition for trusting the prima facie truth of his privileged claim isn't present. Blake blocks this move by discrediting her observational-interpretive evidence in the following way: The fact that Anna has low levels of mindfulness demonstrated by her high reactivity makes her unable to give his alternative explanation its due consideration. Notice that Blake here uses his observational-interpretive evidence of Anna's inadequate mindfulness to block Anna's defeater.

Anna in a second argument points out her observational-interpretive evidence that Blake is selfish and not a good person, and this implies Blake really doesn't love her. Blake blocks this defeater with his observational-interpretive evidence that her speech is influenced by her documented Delusional Disorder disability. Notice that Blake here doesn't use observational-interpretive evidence of her mindfulness level to support his response to her on this second defeater, as he did with the first defeater. Nearly all the different types of observational-interpretive evidence listed in the table "Eight types of observational-interpretive evidence" above are at play and evaluated.

What is very unique and very useful about the Assessment Framework is that it can accommodate all the different varieties of observational-interpretive evidence in peer disagreements about privileged self-knowledge claims. In a sophisticated peer disagreement about a privileged self-knowledge claim we should expect both that all the eight different types of observational-interpretive evidence are at play and that when one uses a particular type of observation evidence as a basis for an argument the other disputant is quick to challenge it with her own four types of observational-interpretive evidence.

The Assessment Framework has been shown to support the strong intuition of this case study to see the claimant should be steadfast in his privileged self-knowledge claim. We have seen that it very nicely accommodates the varieties of arguments based on the different types of observational-interpretive evidence.

Know Yourself in disagreements from the claim-critic's perspective

We will now consider how the claim-critic knows herself in peer disagreements over the claimant's privileged self-knowledge claim. While the claim-critic isn't the one who has the privileged self-knowledge claim, she does learn something about herself in the target disagreement. She learns how well her four different types of observational-interpretive evidence hold up to the four types of observational-interpretive evidence of the claimant.

When the strong intuition is that the claim-critic should conciliate

We first see to what extent the Assessment Framework accounts for, and helps us understand better, a case study where there is a strong intuition that the claim-critic should conciliate. Consider the RACIST case study:

> **RACIST: "I know I am not racist."**
> You are a white faculty member at a large public university, and you are charged with other colleagues to decide a tenure review for another colleague who is African American. You are terrified of making a racist decision, but at the same time you are equally dedicated to upholding standards of teaching and scholarship that you believe are good for all races and for the success of your university. At the preliminary tenure review committee meeting, you express your concerns about this case: "He is close to fulfilling the standards of scholarship that we need to uphold the quality of our PhD program, but he misses the mark." Later that day at a bar you frequent, your friend and tenure committee colleague, Samantha, wants to talk to you about what you said. She helped you skillfully get through many interpersonal and academic problems, and so you consider her an epistemic peer on this issue. She says, "I agree with you that his scholarship isn't the greatest. But, it is a little better than you estimate. I am wondering if you have considered that you may be biased in the estimate of his scholarship?" You say, "If you are implying that I am racist, you are wrong. I know I am not racist. I have many friends that are African American, and I have never discriminated due to race." Samantha says, "I am not saying you are KKK racist, or Jim Crow racist. But, based on my observation, I think you might have a little of what Chris Rock at the Oscars called 'Sorority Racism' where you have black friends and you love the idea that they have equal opportunities, but you don't want them in your inner group." She points out how you avoid African Americans at conference receptions, and how you don't come to the bar when you know a specific African American faculty person is likely to be there. You listen carefully to her, and for every instance of avoidance that she cites, you find extenuating circumstances for your behavior. You point out how there is one African American colleague at conferences you always avoid because that person doesn't work on the issue you write on, and at conferences you mostly only want to talk to people writing on your topic. And you avoid the African American colleague at the bar because she always wants to talk about sports, and you hate talking about sports. You then point out how Samantha must not have seen how you like talking to particular African Americans at conferences and this bar you frequent. You can't see anything in your speech or behavior that indicates you don't want these African American colleagues in your "sorority." What you say makes good sense to Samantha. In fact, she then recalls one time when you were extensively and naturally talking shop with a different African American colleague at a conference. She conciliates, "I am sorry I thought you were a sorority racist, but I still think you underestimate the person's scholarship."

Here the claim-critic conciliates because the extenuating circumstances you brought up make sense, and because she then remembers observational-interpretive evidence that goes against her initial view that you are a sorority racist. What is decisive here for the claim-critic's conciliation isn't observational-interpretive evidence of signs of mindfulness or lack thereof, though Samantha does notice that you didn't have reactivity to her accusation, that you carefully considered her alternative view and observational-interpretive evidence of your behavior, and that you seemed to extensively review your thoughts and motives associated with avoidance of particular African American colleagues.

What is decisive for the claim-critic is your explanation of the extenuating circumstances and the examples Samantha then recalls of when your behavior at a conference implies you exactly aren't a sorority racist. The consideration of signs of judgmental awareness doesn't raise any red flags. The explanation of not wanting to talk to the African American at the conference because he isn't writing on the same topic, and avoiding an African American colleague at the bar because of not liking to talk about sports checks out and makes sense. They are judgments about the African Americans, but they are not automated judgments based on race.

Notice that the Assessment Framework nicely explains both why the claim-critic initially thinks you are a sorority racist, and also explains why Samantha conciliated. Initially, the claim-critic thought she had observational-interpretive evidence of your behavior that implies you don't have the mental state you think you have, that is, the mental state of not being a sorority racist. After carefully listening to you, your observational-interpretive evidence of your own behavior and speech, recalling a counter example to your initial claim, and, to a lesser degree, the absence of signs of automated judgmental awareness, she found the full observational evidence doesn't imply that you don't have the mental state you think you have. The core conditional statement of the Assessment Framework nicely helps one understand the changing dynamics of this case study exactly because it allows for the different ways in which "signs of judgmental awareness" and "observational-interpretive evidence of behavior and speech" can apply, and how this influences the charge of "implying the privileged self-knowledge claim … is unjustified." The claim-critic concedes because the observational evidence you share supports your view that you are not racist. And the observational-interpretive evidence Samantha presents has been rebutted.

When the strong intuition is that the claim-critic should remain steadfast

We first see to what extent the Assessment Framework accounts for, and helps us understand better, a case study where there is a strong intuition that the claim-critic should remain steadfast. Consider the INTENT case study:

> **INTENT: "I know that I didn't intend to kill my husband!"**
> Chris is a juror in a murder trial that hinges on whether the defendant, Sam, intended to kill her husband. He listened to her testimony:
>
> > I know that I didn't intend to kill my husband! Such intent goes against everything that I am. At a time when he was always at the office and there were a spate of daytime home invasions in our neighborhood, I was taking the gun out of my purse in order to clean it after target practice. Suddenly, I feel and hear someone reaching around my shoulder trying to grab my hand. [*Now crying and pausing to recover enough composure to talk*] He must not have seen the snub nose revolver in my hand. When he pulled my hand towards him, the gun went

> off. I didn't know at that point that it was my husband. I feared for my life that I was being attacked. As he laid dying, I saw it was my husband, and he said, "I just wanted to dance with you like we always did when we were doing well."
>
> There is a good amount of evidence contradicting her testimony: They were in the middle of a terrible custody battle for their children; there is no gun residue on her husband's hand, which is expected when near a gun firing; the manager at the range testified that he complimented her about how well she was shooting the target, to which she replied, "If only it were my husband;" two of her colleagues recall her saying a day before the incident in the break room after crying about the difficulties with her husband, "I should just kill him;" the blood splatter, position of his body when he fell, and the entry point of the bullet were all highly inconsistent with her testimony. And there were a few scratch marks on her hand and legs as if a fight happened. An expert crime scene investigator testified that he must have been shot from at least six feet away. Given all this evidence, Chris believes she probably murdered her husband, but he also thinks there is reasonable doubt. Perhaps her shirt was between his hand and the gun such that he didn't get gun residue. And he just thinks there is no way that she could have faked the extremely deeply felt and sincere testimony she gave. Then, a well-respected forensic psychologist who specializes in defendant testimony made in extremely stressful situations takes the stand and describes many critically acclaimed studies she and her colleagues have published that statistically prove something about people who, on one hand, are in extreme psychologically stressful situations like Sam, and, on the other, do something in that situation that goes against their self-concept. Most of the time they sincerely believe they didn't intend to do it even when they actually did intend to do it at the time. The psychologist also points to tell-tale signs that Sam is repressing the memories that threaten her understanding of herself. She forcefully denies having said, "If only it were my husband," and "I should just kill him" even when there appears to be no bias in those who testified to the remarks. The forensic psychologist says based on the evidence they have found and the relevant psychological testing they have had her do, she is suffering from Brief Psychotic Disorder with a stressor (DSM-5 298.8). Because of all this additional evidence Chris now believes Sam intended to murder her husband.

Notice that before the forensic psychologist took the stand, Chris thought there was very strong evidence that she intended murder; but Chris believes it wasn't enough for a conviction of murder. What is enough, though, is the evidence from empirical studies that the forensic psychologist presents along with signs of judgmental awareness. The former evidence brings him closer to a belief that they should convict, but he has some doubts that empirical research of people in stressful situations can reflect on the extreme stress Sam would have experienced. But the extreme reactivity she has to the testimonies of the seemingly unbiased people reporting what she said just puts him over the line such that he is convinced they should convict Sam. The manager at the range only interacted with her that one time, and the colleagues were new hires without enough time there to dislike her so much that they would lie about what she said. The extreme denials that she said, "If only it were my husband," and "I should just kill him" are so emotional. He reasons that if she were in her right mind and she wanted to beat the charge of murder, she would just say she was joking about the things she said. The observational-interpretive evidence of signs of judgmental awareness fit the explanation of the forensic psychologist that she wasn't in her right mind due to the extreme stress. The evidence of judgmental awareness was the breaking point for Chris to be confident in recommending conviction.

The Assessment Framework can yield a good understanding of what happened with Chris to make him in the end recommend conviction. As the Assessment Framework specifies, a necessary condition for trusting a person's privileged self-knowledge claim is that there is no observational-interpretive evidence that implies the privileged self-knowledge claim is unjustified. Under normal situations, the weight of the evidence would be

adequate to imply the privileged self-knowledge claim is unjustified. But this is a murder case, and for the conflicting observational-interpretive evidence to count as implying the privileged self-knowledge claim is unjustified, there would have to be no reasonable doubt about the merits of the conflicting observational-interpretive evidence; and Chris felt there was such reasonable doubt. But then the expert witness of the forensic psychologist and the signs of judgmental awareness in the extreme reactivity fit so well with the other observational-interpretive evidence that the observational-interpretive evidence now decisively implies the privileged self-knowledge claim is false.

What these four case studies tell us

In each one of these diverse case studies of target disagreements, the prime facie truth status of a privileged self-knowledge claim is contested. In each the core conditional statement at the heart of the Assessment Framework helps us naturally and organically see what is going on at the deepest level of the target disagreements. In each a necessary condition must be fulfilled in order for the prima facie truth status to be trusted. In each of the four very different target disagreements, the Assessment Framework helps us get to the heart of the matter. These findings suggest that the Assessment Framework is useful and reflects the way people actually engage target disagreements. In each of the case studies the four types of observational-interpretive evidence of each disputant are taken into account.

CONCLUSION

Know Yourself Better Through Peer Disagreements

While the derivation and successful application of this Assessment Framework is a major goal of this dissertation, we also learn something deeper about self-knowledge through peer disagreements which is implied by the Assessment Framework and all the work done to derive it. I take there to be a distinction between Knowing Yourself in versus through target disagreements. The former specifies a particular context in which specific self-knowledge is disputed and assessed; and up to this point in this dissertation we have been describing "Knowing Yourself" in specific contexts. But through the process of thinking about the components of target disagreements and the Assessment Framework, there are four things we can say about the epistemic state of any privileged self-knowledge claim that apply to the context of peer disagreements just as much as they apply to any other context, one about the need for corroborating privileged self-knowledge claims, a second about the need for observational-interpretive feedback from others, a third about mindfulness, and another about the value of target disagreements for any attempt to Know Yourself.

One way that you –I use the word "you" here since the Delphic oracle uses the second person singular pronoun—can Know Yourself (γνῶθι σεαυτόν) better through target disagreements is by seeing that you often need to corroborate your privileged self-knowledge claims, and for this you need all the types of observational-interpretive evidence from both yourself and from your claim-critic. Of course, it is appropriate normally to trust your own privileged self-knowledge claims prima facie. But, situations aren't always normal. Here are some examples of situations that aren't normal and, so, require corroboration:

Situations that aren't normal and, so, require corroboration

1) The stakes for knowing yourself are high, like when:
    a) You need to know if you love someone such that you would be happy marrying the person.
    b) You need to know if you would be happy choosing a future possible career.
    c) You suspect that you might be racist.
    d) You suspect that you might be jealous of a person at work.
    e) You must know for sure that you are scolding your child for the reasons you think you are.

2) You have doubts about whether you have the mental state you think you have.
3) You want to be proactive assuring that your claim to privileged self-knowledge reflects reality.
4) You want to be sure that you actually believe something you think you believe.
5) A strong peer contests your privileged self-knowledge claim.

Here is one example of a high stakes situation where one would need assurance that the mental state claimed is actually the mental state at play.

Abortion Assurance

A woman, Julie, is considering having an abortion. She believes that abortion is the best path to take for all parties, and she even tells her friends that she will have it done. But, she also knows that this decision is a big one for her. Julie just recently came to see abortion in a positive way, whereas in the past she thought it was horrible to have an abortion. Even though she is very confident this is the right decision for her that will make her happier in the future, she wants more assurance that she will be okay with this decision. She has found it very helpful in the past to talk to a therapist about life decisions. When she has described what she is thinking about an issue and the reactions she has to such thoughts, therapists in the past have helped her see relevant things in her thoughts and behaviors, whether supporting her life decision or not. And consequently, she has made decisions with even more confidence. So, she calls to make an appointment with a talk therapist she has worked with in the past.

Notice here that Julie seeks to corroborate her belief that she should get an abortion by considering the evidence that her therapist gathers from observing her speech and behavior, and also evidence of judgmental awareness. She weighs the observational-interpretive evidence her therapist has of her against the observational-interpretive evidence she has of her own behavior and signs of judgmental awareness. The therapist looks for signs of reactivity when Julie describes her situation, since this is a clear sign of judgmental awareness which is evidence that a belief is less likely to be correct. Her happiness and mental health in the future will be influenced by this decision, and this is why she is seeking more assurance than is afforded by the prima facie status of her privileged self-knowledge claim.

Notice also that Julie could get the added assurance she is looking for without having a disagreement with her therapist. The therapist could just give her supporting evidence from her behavior and speech that she would be, with respect to her mental health, okay having an abortion. While a target disagreement isn't a necessary condition for a demand for corroboration, we will see how it is a sufficient one for such a demand, and a potentially groundbreaking one at that.

The list above shows just some of the ways that there can be a need for more assurance than the Prima Facie Norm can afford. The Prima Facie Norm is just that, a norm; and it says, "Trust a disputant's privileged self-knowledge claim prima facie if there are no or little signs that it is unjustified." Even when normally there

aren't good reasons to reconsider whether your claimed mental state is what you think it is, in the atypical situations above, there are good reasons to seek more assurance that you have the mental state.

Of course, one can make a privileged self-knowledge claim without at the same time being careful to monitor for signs of judgmental awareness and observational-interpretive evidence to the contrary. But, if one wants to ensure that one has the mental state claimed, then one must corroborate one's claim against the observational-interpretive evidence. If one wants more preemptive assurance that what you say is your mental state really is your mental state, one must look for signs of judgmental awareness and observational-interpretive evidence that imply one doesn't really have the mental states one claims one has.

In such atypical situations it makes sense to seek more assurance, since then there is a need to rule out more of the many ways that attempts at self-knowledge are foiled by psychological, physical, and cultural barriers. People can be wrong about even what seems to be the most secure self-knowledge claim such as "I know I love you" or "I have a headache". Chapter Four showed that even the most diligent scholars of peer disagreement sometimes don't adequately recognize the fragile nature of privileged self-knowledge claims. Even in the best of conditions, one can be mistaken about what one's mentals states and attitudes are.

What this means, consequently, is that, while one can normally trust one's own, and those of others', privileged self-knowledge claims, one must be ever vigilant looking for signs of judgmental awareness and observational-interpretive evidence that imply one's privileged self-knowledge claim is unjustified. Julie, in the case study above, is proactive in going to the therapist. In doing so, she takes the risk of finding observational-interpretive evidence that goes against the belief that she should have the abortion. While that would be unsettling, she would possibly avoid more extreme dysphoria in the future as a result of her vigilance.

A necessary condition for such corroboration is that there is an essential interrelatedness between knowing one's mental states through privileged access and knowing one's mental states through observational-interpretive access. The interrelatedness comes from the fact that people usually behave and speak in accord with their mental states and attitudes. That means that if one's behaviors and speech aren't in accord with the mental state, this is evidence that the person doesn't have the mental state. Julie knows this very well, at least

intuitively, because she is going to her therapist to see if she finds observational-interpretive evidence in her behaviors and speech that imply she doesn't have the mental state she believes she has.

And it isn't that one's actions have to line up perfectly with the mental state one has. Think about the LOVE case study where the husband says and does some mean things to his wife which many would take as implying that he really doesn't have the love for his wife that he thinks he has. But Blake had compelling extenuating circumstances that preserved the legitimacy of his claim to love her.

From here on I would like to talk about what we learned <u>through</u> peer disagreements in the way that the Delphic maxim addresses us even today, in the second person, singular, aorist, active, imperative of Know Thyself (γνῶθι σεαυτόν). I do this because it represents more the perspective of the actual peer disagreement about privileged self-knowledge claims where the issue is very personal and where one addresses the other as "you."

You also Know Yourself better though target disagreements by seeing the extent to which mindfulness influences your understanding of yourself and others. Research proves mindfulness improves the reliability of privileged self-knowledge claims. It also helps you evaluate your own such claims with your own observational-interpretive evidence of your behavior and speech. And it helps you weigh the observational-interpretive evidence brought by your claim-critic. With these crucial benefits mindfulness influences your understanding of yourself and others.

We learn another thing about the epistemic state of attempts at self-knowledge, namely, that you need the feedback of others if you are to be proactive about you claims to self-knowledge based on privileged access, or if you are in any of the other atypical conditions listed above. Recall that when there is a target disagreement with your own privileged self-knowledge claim at issue, you have four types of observational-interpretive evidence for corroborating your privileged self-knowledge claim. And you also potentially have the observational-interpretive evidence of your claim-critic who asserts that her observational-interpretive evidence of your behaviors and speech imply that you don't actually have the mental state you think you have. Of course you, as the claimant, weigh your observational-interpretive evidence against that of your claim-critic's.

But when you are weighing your observational-interpretive evidence against that of your claim-critic, consider this: You have barriers to self-knowledge your claim-critic doesn't have. You have, as does everyone, strong and unconscious ulterior motives whose only job is to make sure that you feel good about yourself, like the self-enhancement motive. Proof of such ulterior motives is the Dunning/Kruger effect which demonstrates that people usually inflate their own skills and achievements over those of others (Dunning, 2003, 2004, 2005). There are many other psychological, physical, and cultural barriers to self-knowledge you have which your claim-critic doesn't have. You want to be right about what you claim to know about yourself through privileged access. And you are highly motivated to think you are right about such things whether you are actually right about them or not. You have much invested in being right about your claims to privileged self-knowledge. Of course, your claim-critic can have much invested in being right about you not actually having the self-knowledge you claim you have, perhaps even so much invested that she would believe you are wrong about your claim even when you really are right about it.

But you have a possible conflict of interest that your claim-critic can't have, namely, you are invested in a claim to know something that makes you who you are, a particular mental state, while the claim-critic is invested only in a belief that you don't have the mental state you think you have. If you come to see that your claim-critic is a strong peer, and you come to think the person has no signs of judgmental awareness whereby their observational-interpretive evidence about your behaviors and speech are tainted by their own biases, that observational-interpretive evidence of your behaviors and speech from the claim-critic can be very valuable for you; it can help you more likely form true privileged self-knowledge claim rather than false one. What I have just described here would be represented in a situation where you have a good and impartial therapist, friend, parent, sibling, or mentor who gives you feedback about her observations of your behaviors and speech. And you may even have a target disagreement with that person. You Know Yourself better <u>through</u> these types of peer disagreements because you understand that peers can have observations evidence about yourself that you can't always get for yourself, observational-interpretive evidence that helps you Know Yourself better.

A fourth thing you can know about the epistemic state of privileged self-knowledge claims follows from all that has been said, namely, you should often see target disagreements as opportunities. There is an epistemic

value of target disagreements. Through target disagreements the claim-critic can help you see things about yourself that you can't often see, as has been discussed. When you recognize that any privileged self-knowledge claim is fragile, that you have a tendency towards judgmental awareness which creates barriers to self-knowledge claims reflecting the reality of your mental states, and that other peers can help us spot judgmental awareness, peer disagreements about privileged self-knowledge claims takes on a new epistemic value.

I believe Socrates would approve of this approach to self-knowledge through target disagreements. Socrates' method for understanding ethical or other concepts is disagreement oriented. It is through some of the most challenging disagreements that one comes to a better understanding of the topic at hand. His Socratic method works by detecting judgmental awareness in the sense that he points out unsubstantiated assumptions and automated judgments. So, in the conclusion of this dissertation the Delphic maxim used also by Socrates and Plato is extended to come to the following deeper understanding of the epistemic human condition as regards self-knowledge: Know Yourself (γνῶθι σεαυτόν) better through target disagreements with mindfulness both in privileged and observational-interpretive access integrated.

WORKS CITED

Abbasi, Maryam, Nima Ghorbani1, Amir Hossein Imani, and Sahar Tahbaz Hoseinzadeh. 2020. “Exploring the Mediating Role of Integrative Self-Knowledge in the Relationship between Mindfulness and Well-Being in the Context of a Mindfulness-Based Stress Reduction Program.” *International Journal of Psychology* 56, no. 2.

Alicke, Mark D; Guenther, Corey L; Zell, Ethan. 2012. "Social self-analysis: Constructing and maintaining personal identity." *In Handbook of Self and Identity*, edited by Mark Leary, 291-308. New York: Guilford Press.

Anderson, Rebecca Cogwell. 2015. "Fooling ourselves and not even knowing it [Review]." PsycCRITIQUES 60 (20).

Anseel, Frederik, and Filip Lievens. 2006. "Certainty as a moderator of feedback reactions? A test of the strength of the self-verification motive." *Journal of Occupational and Organizational Psychology* 79 (4):533-551.

Arch, Joanna. J, and Michelle G Craske. 2006. "Mechanisms of mindfulness: Emotion regulation following a focused breathing induction." *Behavior Research and Therapy* 44:1849-1858.

Baars, B. *A.* 1988. *Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baer, Ruth. 2003. "Mindfulness training as a clinical Intervention: A conceptual and empirical review." *Clinical Psychology: Science and Practice* 10:125-143.

Baer, Ruth A., Gregory T. Smith, Jaclyn Hopkins, Jennifer Krietemeyer, and Leslie Toney. 2006. “Using Self-Report Assessment Methods to Explore Facets of Mindfulness.” *Assessment (Odessa, Fla.)* 13 (1): 27–45. https://doi.org/10.1177/1073191105283504.

Bar-On, Dorit. 2004. *Speaking My Mind: Expression and Self-Knowledge*. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Clarendon Press.

Batory, Anna Maria. 2015. "What self-aspects appear significant when identity is in danger? Motives crucial under identity threat." *Journal of Constructivist Psychology* 28 (2):166-180.

Bauer, Isabelle M., and Roy F. Baumeister. 2013. "Self-knowledge." In *The Oxford Handbook of Cognitive Psychology*, 905-917. New York, NY: Oxford University Press; US.

Bem, Daryl. 1970. "Testing the self-perception explanation of dissonance phenomena." *Journal of Personality and Social Psychology* 14 (1):23-3.

Bem, Daryl. 1972. "Self-perception theory." *Advances in Experimental Social Psychology* 6:1-62.

Bergmann, Michael. 2009. “Rational Disagreement after Full Disclosure.” *Episteme: A Journal of Individual and Social Epistemology* 6, no. 3: 336–53.

Berger, J. D, and Herringer, L. G. (1991). Individual differences in eyewitness recall accuracy. *Journal of Social Psychology*, 131, 807-813.

Brewer, Judson A., Patrick D. Worhunsky, Jeremy R. Gray, Yi-Yuan Tang, Jochen Weber, and Hedy Kober. 2011. "Meditation Experience is associated with differences in default mode network activity and connectivity." Proceedings of the National Academy of Sciences 108:20254–20259.

Bothwell, R. K., Deffenbacher, K. A., and Brigham, J. C. 1987. Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, 72, 691-695.

Britton, Willoughby B., Ben Shahar, Ohad Szepsenwol, and W.Jake Jacobs. 2012. "Mindfulness-based cognitive therapy improves emotional reactivity to social stress; Results from a randomized controlled trial." *Behavior Therapy* 43:365-380.

Brooks, Alison. 2014. "Get excited: reappraising pre-performance anxiety as excitement." *Journal of Experimental Psychology: Applied* 143:1144–1158.

Brown, K.W., and Ryan, R. M. (2003). The benefits ofbeing present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, 84(4), 822–848.

Brown, Kirk Warren, Netta Weinstein, and J. David Creswell. 2012. "Trait mindfulness modulates neuroendocrine and affective responses to social evaluative threat." *Psychoneuroendocrinology* 37:2037-2041.

Brown, K.W., and Ryan, R. M. 2003. The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, 84(4), 822–848.
Campbell, Jennifer D. 1990. "Self-esteem and clarity of the self-concept." *Journal of Personality and Social Psychology* 59:538-549.
Carlson, Erika N. 2013. "Overcoming the barriers to self-knowledge: Mindfulness as a path to seeing yourself as you really are." *Perspectives on Psychological Science* 8 (2):173-186.
Carlson, Erika, Simine Varize, and Thomas Oltmanns. 2013. "Self-other knowledge asymmetries in personality pathology." *Journal of Personality* 81 (2):155-170.
Carruthers, Peter. 2007. "The illusion of conscious will." *Synthese* 159 (2):197-213.
Carruthers, Peter. 2011. *The Opacity of Mind: The Integrative Theory of Self-Knowledge*. Oxford: Oxford.
Cassam, Quassim. 2014. *Self-Knowledge for Humans*. Oxford University Press.
Chaiken, Shelly; Baldwin, Mark. 1981. "Affective-Cognitive Consistency and the Effect of Salient Behavioral Information on the Self-Perception of Attitudes Shelly Chaiken University of Toronto Ontario, Canada Mark W. Baldwin University of Waterloo." 1981. *Journal of Personality and Social Psychology* 41 (1): 1–12.
Chaiken, Shelly; Baldwin, Mark. 2008. "Affective-cognitive consistency and the effect of salient behavioral information on the self-perception of attitudes." In *Attitudes: Their structure, function, and consequences*, edited by Fazio, Russel; Pretty, Richard. New York, NY. Psychology Press.
Chambers, Richard, Barbara Chuen Yee Lo, and Nicholas B. Allen. 2008. "The impact of intensive mindfulness training on attentional control, cognitive style, and affect." *Cognitive Therapy and Research* 32:303–322.
Christensen, David. 2018. "On Acting as Judge in One's Own (Epistemic) Case." *Proceedings and Addresses of the American Philosophical Association* 92: 207–35.
Christensen, David. 2009. "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4, no. 5: 756–67.
Christensen, David. 2007. "Epistemology of Disagreement: The Good News." *Philosophical Review* 116, no. 2: 187–217.
Christensen, David. 2011. "Disagreement, Question-Begging, and Epistemic Self-Criticism." *Philosophers' Imprint* 11, no. 6: 1–41.
Christensen, David. 2009. "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4, no. 5 (September): 756–67.
Corcoran, K.M., H. Farb, Anderson A., and Z.V. Segal. 2010. "Mindfulness and emotion regulation: Outcomes and possible mediating mechanisms." In *Emotion Regulation and Psychopathology: A Transdiagnostic Approach to Etiology and Treatment*, edited by A.M. Kring and D.M. Sloan, 339-355. New York: Guilford Press.
Creswell, J. David, Baldwin M. Way, Naomi I. Eisenberger, and Matthew D. Lieberman. 2007. "Neural correlates of dispositional mindfulness during affect labeling." *Psychosomatic Medicine* 69:560-565.
Critcher, Clayton R., and Thomas Gilovich. 2010. "Inferring attitudes from mindwandering." *Personality and Social Psychology Bulletin* 36 (9):1255-1266.
Dummel, Sebastian and Jutta Stahl. 2018. "Mindfulness and the Evaluative Organization of Self Knowledge," *Mindfulness* 10: 352–65.
Dummel, Sebastian. 2018. "Relating Mindfulness to Attitudinal Ambivalence Through Self-Concept Clarity." *Mindfulness* 9 (2018): 1486–93.
Dunning, D. 2005. *Self-insight: Roadblocks and Detours on the Path to Knowing Thyself*. New York, NY: Psychology Press.
Dunning, D, Johnson, K., Ehrlinger, J. and Kruger, J. 2003. 'Why People Fail To Recognize Their Own Incompetence.' *Current Directions in Psychological Science*, 12: 83–6.
Emanuel, Amber S., John A. Updegraff, David Kalmbach, and Jeffrey A. Ciesla. 2010. "The role of mindfulness facets in affective forcasting." *Personality and Individual Differences* 49:815-818.

Farb, Norman A. S., Adam K. Anderson, Helen Mayberg, Jim Bean, Deborah McKeon, and Zindel V. Segal. 2010. "Minding one's emotions: Mindfulness training alters the neural expression of sadness." *Emotion* 10:25-33.

Feldman, Greg, Jeff Greeson, and Joanna Senville. 2010. "Differential effects of mindful breathing, progressive muscle relaxation, and loving-kindness meditation on decentering and negative reactions to repetitive thoughts." *Behaviour Research and Therapy* 48:1002-1011.

Gopnik, Alison. 1993. "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality." *Behavioral and Brain Sciences* 16: 1–14.

Gray, Marcus, Neil Harrison, Stefan Wiens, and Hugo Critchley. 2007. "Modulation of emotional appraisal by false physiological feedback during fMRI." *PLoS ONE* 2 (6).

Guadagno, Rosanna E., Lankford, Adam, Muscanell, Nicole L., Okdie, Bradley M., and McCallum, Debra M. 2010. "Social influence in the online recruitment of terrorists and terrorist sympathizers: Implications for social psychology research." *Revue Internationale De Psychologie Sociale* 23 (1):25-56.

Hayes, Steven C., Jason B. Luoma, Frank W. Bond, Akihiko Masuda, and Jason Lillis. 2006. "Acceptance and Commitment Therapy: Model, Processes and Outcomes." *Behaviour Research and Therapy* 44 (1): 1–25. https://doi.org/10.1016/j.brat.2005.06.006.

Hanley, Adam W, and Eric L. Garland. 2016. "Clarity of mind: Structural equation modeling of associations between dispositional mindfulness, self-concept clarity and psychological well-being." *Personality and Individual Differences* 106:334-339.

Hanley, Adam. 2017. "Clarity of Mind: Structural Equation Modeling of Associations between Dispositional Mindfulness, Self-Concept Clarity and Psychological Well-Being Adam W. Hanley, PhD ∗, Eric L. Garland, PhD." *Personality and Individual Differences*, no. 106: 334–39.

Hanley, Adam W., and Eric L. Garland. 2017. "The Mindful Personality: A Meta-Analysis from a Cybernetic Perspective." *Mindfulness* 8, no. 6 (2017): 1456–70. https://doi.org/10.1007/s12671-017-0736-8.

Hargus, Emily, Catherine Crane, Thorsten Barnhofer, and J. Mark. G. Williams. 2010. "Effects of mindfulness on meta-awareness and specificity of describing prodromal symptoms in suicidal depression." *Emotion* 10:34-42.

Hill, Christina, and John Updegraff. 2012. "Mindfulness and its relationship to emotional regulation." *Emotion* 12:81-90.

Hölzel, Britta K., Sara W. Lazar, Tim Gard, Zev Schuman-Olivier, David R. Vago, and Ulrich Ott. 2011. "How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective." *Perspectives on Psychological Science* 6 (6):537-559.

Inamori, Yoshio. 1979. "Effects of false heart rate feedback on cognitive appraisal and physiological responses to emotional stimuli." *Japanese Psychological Research* 21 (3):153-157.

Ito, Tiffany A., Krystal W. Chiao, Patricia G. Devine, Tyler S. Lorig, and John T. Cacioppo. 2006. "The Influence of Facial Feedback on Race Bias." *Psychological Science* 17 (3):256-261.

Jha, Amishi P., Jason Krompinger, and Michael J. Baime. 2007. "Mindfulness training modifies subsystems of attention." *Cognitive, Affective, and Behavioral Neuroscience* 7:109–119.

Jha, Amishi P., Elizabeth A. Stanley, Anastasia Kiyonaga, Ling Wong, and Lois Gelfand. 2010. "Examining the protective effects of mindfulness training on working memory capacity and effective experience." *Emotion* 10:54–64.

Kabat-Zinn, Jon. 1994. *Wherever You Go, There You Are*. New York: Hyperion.

Kelly, Thomas. 2013. "Disagreement and the Burdens of Judgment." In *The Epistemology of Disagreement: New Essays*, Ed. David Christensen and Jennifer Lackey. Oxford: Oxford University Press, 31–53," 2013.

Keng, Shian-Ling, Moria J Smoski, Clive J Robins. 2011, *Clin Psychol Review*. Aug;31(6):1041-56. doi: 10.1016/j.cpr.2011.04.006. May 13.

Kernis, Michael H., and Brian M. Goldman. 2006. "A multicomponent conceptualization of authenticity: theory and research." Advances in Experimental Social Psychology 38:283-357.

King, Nathan L. "Disagreement: What's the Problem? Or a Good Peer Is Hard to Find." *Philosophy and Phenomenological Research* 85, no. 2 (September 2012): 249–72.

Koole, Sander L., Olesya Govorun, Clara M. Cheng, and Marcello Gallucci. 2009. "Pulling yourself together: Meditation promotes congruence between implicit and explicit self-esteem." *Journal of Experimental Social Psychology* 45:1220-1226.

Korsgaard, Christine. 2009. "The Activity of Reason" *Proceedings and Addresses of the American Philosophical Association* 83, no. 2: 23–43.

Lackey, Jennifer. 2010. "What Should We Do When We Disagree?" In *Oxford Studies in Epistemology: Volume 3*. Oxford: Oxford Univ Pr, 2010. http://ezproxy.lib.utexas.edu/login?url=https://search.ebscohost.com/login.aspx?direct=truedb=phlAN=PHL2155481site=ehost-live.

Laird, James. 2007. Feelings: *The perception of self*. Series in affective science. New York, NY, Oxford.

Lawlor, Krista. 2009. "Knowing What One Wants." *Philosophy and Phenomenological Research* 79: 47–75.

Levoy, Emily, Asimina Lazaridou, Judson Brewer, and Carl Fulwiler. 2017. "An Exploratory Study of Mindfulness Based Stress Reduction for Emotional Eating." *Appetite* 109 (February): 124–30. https://doi.org/10.1016/j.appet.2016.11.029.

Lueke, Adam, and Bryan Gibson. 2015. "Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding." *Social Psychological and Personality Science* 6 (3):284-291.

Lattimore, Paul. 2020. "Mindfulness-Based Emotional Eating Awareness Training: Taking the Emotional out of Eating." *Studies on Anorexia, Bulimia and Obesity* 25: 649–57.

Ma, S. Helen, and John D. Teasdale. 2004. "Mindfulness-based cognitive therapy for depression: Replication and exploration of differential relapse prevention effects." *Journal of Consulting and Clinical Psychology* 72:31–41.

Makkar, Steve R, and Jessica R Grisham. 2013. "Effects of false feedback on affect, cognition, behavior, and postevent processing: the mediating role of self-focused attention." *Behavior Therapy* 44:111-124.

Modinos, Gemma, Johan Ormel, and Andre´ Aleman. 2010. "Individual differences in dispositional mindfulness and brain activity involved in reappraisal of emotion." *Social Cognitive and Affective Neuroscience* 5:369-377.

Molouki, Sarah, and Emily Pronin. 2015. "Self and Other." In *APA Handbook of Personality and Social Psychology*, Volume 1: Attitudes and Social Cognition., edited by Mario Mikulincer, Phillip R. Shaver, Eugene Borgida, John A. Bargh, Mario Mikulincer, Phillip R. Shaver, Eugene Borgida and John A. Bargh, 387-414.

Morillo-Sarto, Héctor, Yolanda López-del-Hoyo, Adrián Pérez-Aranda, Marta Modrego-Alarcón, Alberto Barceló-Soler, Luis Borao, Marta Puebla-Guedea, Marcelo Demarzo, Javier García-Campayo, and Jesús Montero-Marin. 2023. "'Mindful Eating' for Reducing Emotional Eating in Patients with Overweight or Obesity in Primary Care Settings: A Randomized Controlled Trial." *European Eating Disorders Review* 31 (2): 303–19. https://doi.org/10.1002/erv.2958.

Moran, Richard. 2001. *Authority and Estrangement*. Princeton University Press.

Nisbett, Richard E., and Timothy DeCamp Wilson. 1977. "Telling more than we can know: Verbal reports on mental processes." *Psychological Review* 84:231-259.

Papies, Esther K., Lawrence W. Barsalou, and Ruud Custers. 2012. "Mindful attention prevents mindless impulses." *Social Psychological and Personality Science* 3:291-299.

Palmer, Michael. "Lesson 31: Aorist Active Imperatives." University of North Carolina at Chapel Hill. *Helenistic Greek* (blog), 2023. https://hellenisticgreek.com/31.html.

Paul, Sarah K. 2014. "The Transparency of Mind." *Philosophy Compass* 9 (5):295-319.

Perona-Garcelán, Salvador, José M. García-Montes, Ana M. López-Jiménez, Juan Francisco Rodríguez-Testal, Miguel Ruiz-Veguilla, María Jesús Ductor-Recuerda, María del Mar Benítez-Hernández, M. Ángeles Arias-Velarde, María Teresa Gómez-Gómez, and Marino Pérez-Álvarez. 2014. "Relationship between self-focused attention and mindfulness in people with and without hallucination proneness." *The Spanish Journal of Psychology* 17.

Peters, Uwe. 2014. "Self-knowledge and consciousness of attitudes." *Journal of Consciousness Studies* 21 (1-2):139-155.
Prakash, Ruchika Shaurya, Patrick Whitmoyer, Amelia Aldao, and Brittney Schirda. 2017. "Mindfulness and emotion regulation in older and young adults." *Aging and Mental Health* 21 (1):77-87.
Roediger, Henry, Andrew Desoto, and John Wixted. 2012. "The Curious Complexity between Confidence and Accuracy in Reports from Memory." In *Memory and Law,* L. Nadel and W. P. Sinnott-Armstrong (Eds.), 84–118. Oxford University Press, 2012.
Sala, Margarita, Shruti Shankar Ram, Irina A. Vanzhula, and Cheri A. Levinson. 2020. "Mindfulness and Eating Disorder Psychopathology: A Meta-analysis." *International Journal of Eating Disorders* 53 (6): 834–51. https://doi.org/10.1002/eat.23247.
Sedikides, Constantine. 1993. "Assessment, enhancement, and verification determinants of the self-evaluation process." *Journal of Personality and Social Psychology* 65 (2):317-338.
Sedikides, Constantine. 2007. "Self-enhancement and self-protection: Powerful, pancultural, and functional." *Hellenic Journal of Psychology* 4 (1):1-13
Shoemaker, Sydney. 1994. "Self-Knowledge and 'Inner Sense'." *Philosophy and Phenomenological Research* 54: 249–314.
Sosa, Ernest. "The Epistemology of Disagreement," 2010. In *Social Epistemology*. Adrian Haddock, Alan Millar, and Duncan Pritchard. Published to Oxford Scholarship Online.
Swann, William B. 1992. "Seeking "Truth," finding despair: Some unhappy consequences of a negative self-concept." *Current Directions in Psychological Science* 1:15-18.
Swann, William B. 1997. "The trouble with change: Self-verification and allegiance to the self." *Psychological Assessment* 8:177-180.
Tapper, Katy, and Zoyah Ahmed. 2018. "A Mindfulness-Based Decentering Technique Increases the Cognitive Accessibility of Health and Weight Loss Related Goals." *Frontiers in Psychology* 9: 587.
Vago, David R., and David A. Silbersweig. 2012. "Self-awareness, self-regulation, and self-transcendence (S-ART): A framework for understanding the neurobiological mechanisms of mindfulness." *Frontiers in Human Neuroscience* 6 (296):1–30.
Vago, David R. 2014. "Mapping Modalities of Self-awareness in Mindfulness Practice: A Potential Mechanism for Clarifying Habits of Mind." *Annals of the New York Academy of Sciences* 1307: 28–42.
Vanden Bos, G. R. (Ed.). 2007. "Self-Concept." *APA Dictionary of Psychology.* American Psychological Association.
Valins, Stuart. 1966. "Cognitive effects of false heart-rate feedback." *Journal of Personality and Social Psychology* (4):400-408.
Vazire, Simine. 2010. "Who knows what about a person? The self-other knowledge asymmetry (SOKA) model." *Journal of Personality and Social Psychology* 98 (2):281-300.
Vazire, Simine, and Erika Carlson. 2010. "Self-knowledge of personality: Do people know themselves?" *Social and Personality Psychology Compass* 4 (8):605-620.
Vazire, Simine, and Erika N. Carlson. 2011. "Others sometimes know us better than we know ourselves." *Current Directions in Psychological Science* 20 (2):104-108.
Wadlinger, Heather A, and Derek M. Isaacowitz. 2011. "Fixing our focus: Training attention to regulate emotion." *Personality and Social Psychology Review* 15:75-102.
Way, Baldwin M., J. David Creswell, Naomi I. Eisenberger, and Matthew D. Lieberman. 2010. "Dispositional mindfulness and depressive symptomatology: Correlations with limbic and self-referential neural activity during rest." *Emotion* 10:12-24.
Wells, G, and R Petty. 2008. "The effects of overt head movements on persuasion." *Basic and Applied Social Psychology* 1:219-230.
Williams, J. Mark G., John D. Teasdale, Zindel V. Segal, and Judith Soulsby. 2000. "Mindfulness-based cognitive therapy reduces overgeneral auto-biographical memory in formerly depressed patients." *Journal of Abnormal Psychology* 109:150-155.
Wilson, Timothy. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge: Belknap Press.

Wilson, Timothy; Dunn, Elizabeth. 2004. “Self-Knowledge: Its Limits, Value, and Potential for Improvement.” *Annu Rev Psychol*. 55: 493–518.
Wilson, Timothy D. 2009. "Know Yourself." *Perspectives on Psychological Science* 4 (4):384-389.
Wilson, Timothy D., and Elizabeth W. Dunn. 2004. "Self-knowledge: its limits, value, and potential for improvement." *Annual Review of Psychology* 55:493-518.
Wilson, Timothy D., and Daniel T. Gilbert. 2005. "Affective forecasting: Knowing what to want." *Current Directions in Psychological Science* 14:131-134.
Wright, Crispin. 2015. “Self-Knowledge: The Reality of Privileged Access. In: Goldberg S.” In *Externalism, Self-Knowledge and Scepticism: New Essays, Goldberge, S (Ed.)*, 49–74. Cambridge: Cambridge University Press.
Zeidan, Fadel, Susan K. Johnson, Bruce J. Diamond, Zhanna David, and Paula Goolkasian. 2010. "Mindfulness meditation improves cognition: Evidence of brief mental training." *Consciousness and Cognition: An International Journal* 19:597–605.
Hsu, Ti, and Catherine A. Forestell. 2021. “Mindfulness, Depression, and Emotional Eating: The Moderating Role of Nonjudging of Inner Experience.” *Appetite* 160 (May): 105089. https://doi.org/10.1016/j.appet.2020.105089.
Verrier, Diarmuid, and Catherine Day. 2021. “The Moderating Effects of Mindfulness Facets on Psychological Distress and Emotional Eating Behaviour.” *Health Psychology Report* 10 (2): 103–10. https://doi.org/10.5114/hpr.2021.109921.