THE UNIVERSITY OF CHICAGO

ETHICAL DEVELOPMENT AND THE VARIETIES OF SELF-KNOWLEDGE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE HUMANITIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHILOSOPHY

BY
SANTIAGO MEJIA-RIVAS

CHICAGO, ILLINOIS
AUGUST 2016

Para Catalina

Para Ganesh

Para Guadalupe

Por todo ... Y por tanto ...

Table of Contents

List of Figures

Acknowledgments

So many people to thank...

So many thanks to give...

So much gratitude to acknowledge...

Writing a dissertation is sometimes characterized as a survival test. If it is, many individuals were instrumental in my getting alive to the other shore. Without the help of faculty members who interacted with me, the support of students in the philosophy department and the companionship of friends who were not affiliated with the program I would have surely drowned.

I have very deep gratitude to the members of my committee: Martha Nussbaum, the chair, and David Finkelstein and Nancy Sherman, the other two members of the committee. If I am a better thinker, writer and reader, it is in good measure due to having worked with such a great committee. Martha, an exemplar of a far reaching scholar, was also a paradigm of an outstanding mentor. Her firm hand and warm encouragement provided the right environment to face and overcome the difficulties of the project. David was extremely generous with his time, often spending hours going carefully through my work, providing me with sharp criticisms that would significantly improve it. The invaluable writing lessons that I took out of those meetings made the frustrations of scheduling them more than worth it. Nancy's bibliographic suggestions and her empirically informed remarks about my work were always spot on. I will be forever thankful for the time and dedication that the three of

you devoted to my growth as a scholar.

I am also extremely grateful to other faculty members who provided me with support and guidance in different stages of the program. Agnes Callard was always willing to read my work. Her acute questions helped me think about my project in deeper ways. I consider her to be a honorary member of the committee. Brian Leiter helped me to come up with the earliest version of my project and provided me with sound professional suggestions at different stages in my graduate education. Anyone familiar with the work of Jonathan Lear will recognize the pervasive influence that his work has on mine. He also taught me the importance of pursuing ideas you care about despite the fads in the profession. Teaching with Susan Gzsech was great and it is hard to convey how much I enjoyed it. Elizabeth Chandler and Bill Rando pointed me to very fertile ways to think about teaching. By fostering the collaborative and interdisciplinary environment of the Center for Teaching and Learning they allowed me to be nurtured by the wisdom of the great group of scholars who work in it. If I am a better teacher I owe it more than anything to them. Valerie Wallace was an administrator who helped me see what true virtue consists in. I hope I can become half as nice as she is.

Stina Bäckström, Noah Chafets, Tupac Cruz, Dhananjay Jagannathan and Daniel Rodriguez read and criticized a lot of my work and provided me with critical emotional support. Stina read pretty much everything I wrote in my formative years. Her feedback helped me see what it was to be a scholar in philosophy. She also gave me much needed hope and encouragement in my darkest hours in the program. Noah read a huge amount of my work. By indicating how everything I wrote was relevant to Plato he showed me new ways to approach ancient philosophy. His patience in going carefully through my work, proofreading it with utmost care, really helped me polish my own writing skills. I am forever grateful. Pac enlightened our life, cheering our days when he came over to our house. His unique feedback on my work opened up insightful and original directions that I would not have otherwise

pursued. I feel privileged to have lived together with him in the same city so many times and for so many years. Daniel was the kind of guy who is there when you really need him. And it was a relief to be able to share our background and speak about philosophy in our native language. In the last couple of years I was fortunate to interact with Dhananjay and to learn, through these interactions, how to become a more professional philosopher in North America. Thank you all!

I am not sure that I would have made it were it not for my Colombian friends living in Chicago. More than friends they were my family in the U.S. Ana Maria Arjona, Andreas Feldman, Eudald Lerga, Helena Olea, Carlos Manrique, Maria del Rosario Acosta, Pietro Bonaldi and Tupac provided a thread back to the homeland and helped me in the process of sorting out my identity and place as a Latin American Immigrant in the United States. I am also profoundly grateful to my longstanding Colombian friends from school and my early days in college: Sergio Barrero, Gilberto Ordoñez, Mauricio Gonzales and Alejandro Martin. Being so different from one another and so different from me has always given me a healthy perspective on what a flourishing life may look like. I also want to thank my extended family back home. My dad, at least in spirit, supported my journey. And my mom, together with mis tios y tias and primos y primas always supported my protracted life as a student with encouragement and humor. In my dark hours, knowing that all of you are out there makes me feel less lonely. Thank you.

Many other students were instrumental in the development of this project and of myself as a scholar. My cohort was really great and being part of it made my transition into graduate school less traumatic. Stina, Mark Hopwood, Alptekin Sanli, Stephen Shortt and Matt Teichman became good friends. I hope our paths cross many times in the future. At different times during my graduate education, I had valuable interactions with many other students in the program. And even though I am grateful to the whole community of students in the department, I'd like to mention specifically Amichai Amit, Nir Ben Moshe,

Pascal Brixel, Molly Brown, Amos Browne, David Holiday, Katie Howe, Jenny Johnson, Nic Koziolek, Nethanel Lipshitz, Dasha Polzic, Francey Russel, Sophia Sklaviadis, Will Small, Andy Werner and Nate Zuckerman. I would also like to thank a few undergraduates that contributed to the project at different stages: Maximilian Chaoulideer, Peter Goldberg, Tyler Neenan and Roxanne Tam. I am deeply grateful to Greg Rizzolo for having listened to me with patience and compassion, for teaching me valuable practical lessons about the practices of psychotherapy and for helping me recognize that, even if one might be aided by many, one's well being is ultimately one's own burden. Finally, I have profound gratitude to Fiona Lazar and Chloe Pelletier for making my family's life so much brighter and providing the youngsters with alternative models of who one can become. Thanks so much.

Y, claro, gracias infinitas para los tres que viven conmigo. Todo habría sido distinto sin ustedes. Todo habría sido distinto pero no querría yo que fuese distinto. Gracias por aguantar mis inaguantables singularidades, por incluso a veces quererlas. Los otros me ayudaron a nadar pero fueron ustedes los que me dieron el aire para flotar. El peregrinaje es a veces difícil, pero si he avanzado es porque ustedes me han mostrado el camino. Los quiero mucho.

In writing these acknowledgments, in looking back at all the people who have helped me along the way, it dawns on me that writing my dissertation was not a survival test but a growing journey filled with many great companions. Thank you all for making it so.

Abstract

Within virtue ethics, there has been a growing interest in topics surrounding moral education and ethical development. It is often taken to be a truism that self-examination and self-knowledge play an important part in ethical development. This perception might explain why little attention has been paid to problematizing and understanding the role that self-examination and self-knowledge play in such development.

This alleged truism, namely, that self-examination and self-knowledge are important for ethical development has come under attack in recent years from scholars inspired by research in social psychology and cognitive science. Research on depressive realism, for instance, has been used to claim that a certain dose of self-deception is necessary for a flourishing life. Research on associative learning, priming influences, and automatic processes has shown that our behavior is shaped by such a myriad of unconscious influences that the project of knowing oneself appears futile. Even within virtue ethics there has been a slowly growing distrust about the relevance of self-examination and self-knowledge for the development of virtue. It has been objected that the self-centered nature of self-examination is an impediment to the development of virtue, which is outward-oriented. It has also been suggested that in making self-knowledge central for ethical development one ends up portraying such development as excessively intellectual, failing to accommodate virtuous agents who are not very good at knowing themselves.

In the dissertation I respond to these challenges. I show that coming to know yourself, in particular coming to know the unconscious mental states that inform your life, is an essential

part of ethical development. A central part of the argument consists in showing that there is an internal connection between self-examination, self-knowledge and rational agency. In showing this I make perspicuous why self-knowledge should be taken to be, not just one of the many tools that allow us to develop ethically, but a capacity that underwrites any such tool.

Examining the relationship between self-examination, self-knowledge and ethical development is particularly pressing for moral philosophers whose work is, like the work of most virtue ethicists, informed by historical figures. These philosophers have been influenced by thinkers who, for the most part, wrote before the 20^{th}century, a century that saw significant contributions from psychotherapists, cognitive scientists and social psychologists to our understanding of the relationship between self-examination, self-knowledge and ethical development. Engaging with the literature on psychotherapy and psychology allows me to offer a textured account of ethical development that is empirically informed.

At the same time that the dissertation establishes that self-examination and self-knowledge play an important role in ethical development, it investigates a distinction that has been central to epistemology but whose ethical significance has been overlooked: the distinction between what one might call first- and third-personal self-knowledge. In highlighting the ethical significance of this distinction the dissertation puts in contact the work that epistemologists have done on self-knowledge and first-person authority with the work that ethicists have done in moral education and ethical development. Thus, although the dissertation is mainly a contribution to moral philosophy it also sheds some light on epistemology.

The dissertation defends the thesis that knowing oneself first-personally, as opposed to third-personally, is essential for developing virtue. In arguing for this thesis I am attempting to do justice to an insight that is central to most forms of psychotherapy but which has been neglected within virtue ethics (as well as within social psychology and cognitive science). The work of David Finkelstein and Richard Moran provides the framework within which

I argue, against a number of social psychologists and cognitive scientists working on the topic of self-knowledge, that if the aspiration to know oneself is part of a person's ethical development, then it cannot be merely an aspiration to acquire information about oneself but also, and quite importantly, to relate and engage with this information in a first-personal way. I claim that it is only with this kind of self-knowledge that one can aspire to unify rationally the competing elements in one's life into a coherent whole.

I also appeal to the work by Jonathan Lear to elucidate further the nature of first-personal self-knowledge. Moran and Lear agree on the thesis that a certain kind of first-personal self-knowledge is important for ethical development but disagree on how to characterize this kind of self-knowledge. Lear defends an expressive model of first-personal self-knowledge according to which a person has first-personal self-knowledge of her mental state in virtue of the fact that she can *express* it in a self-ascription. Moran defends an agential model of first-personal self-knowledge according to which a person has first-personal self-knowledge of her mental state because she can make up her mind about it deliberately. Although their positions seem in competition with each other I argue that they are actually complementary. Each of these types of first-personal self-knowledge plays a distinctive role in our ethical development. Expressive self-knowledge is required to make explicit patterns of thought that interfere with the person's ability to live well. Agential self-knowledge is the ultimate aim of this kind of self-transformation —a properly human life is one where the person's mental states are formed and can be transformed though her rational assessments of the merits of the situation. These two ways of knowing oneself are connected by the fact that, if things are working well, the former leads to the latter. In fact, I argue that the expressive dimension of self-knowledge facilitates the process whereby instinctive and automatic responses that are not rational become unified with, and transformed by, our self-conscious reflection, a process that ultimately leads to the acquisition of the kind of agential self-knowledge that Moran defends.

I

Introduction

Within virtue ethics, there has been growing interest in topics surrounding moral education and ethical development. It is often taken to be a truism that self-examination and self-knowledge play an important part in ethical development. The perception that this is obvious might explain why little attention has been paid to problematizing and understanding the role that self-examination and self-knowledge play in such development.

However this alleged truism has come under attack in recent years from scholars inspired by research in social psychology and cognitive science. For instance, research on depressive realism, i.e. the thesis that accurate self-knowledge leads to depression, has been used to claim that a certain dose of self-deception is necessary to live a flourishing life. Research on associative learning, priming influences, and automatic processes has shown that our behavior is shaped by such a myriad of unconscious influences that the project of knowing oneself appears futile. Even within virtue ethics there has been a slowly growing distrust about the relevance of self-examination and self-knowledge for the development of virtue. It has been suggested that the self-centered nature of self-examination is an impediment to the development of virtue, which entails attending to the outside world more than —even in preference to— oneself. It has also been suggested that in making self-knowledge central

for ethical development one ends up portraying such development as excessively intellectual, failing to accommodate agents who are virtuous but who are not very good at knowing themselves.

In the dissertation I respond to these challenges. I show that coming to know yourself, in particular coming to know the unconscious mental states that inform your life, is an essential part of ethical development. A central component of the argument consists in showing that there is an internal connection between self-examination, self-knowledge and rational agency. If rationality is essential to a flourishing life, as I will show that it is, then this shows why self-knowledge should be taken to be essential for flourishing.

Examining the relationship between self-examination, self-knowledge and ethical development is particularly pressing for moral philosophers whose work is, like the work of most virtue ethicists, informed by historical figures. These philosophers have been influenced by thinkers who, for the most part, wrote before the 20^{th}century, a century that saw significant contributions from psychotherapists, cognitive scientists and social psychologists to our understanding of the relationship between self-examination, self-knowledge and ethical development. Engaging with the literature on psychotherapy and psychology allows me to offer a textured account of ethical development that is empirically informed, but which has been seldom taken into account by virtue ethicists.

At the same time that the dissertation establishes that self-examination and self-knowledge play an important role in ethical development, it investigates a distinction that has been important in epistemology but whose ethical significance has been overlooked: the distinction between what one might call first- and third-personal self-knowledge. In highlighting the ethical significance of this distinction the dissertation puts the work that epistemologists have done on self-knowledge and first-person authority in contact with the work that ethicists have done in moral education and ethical development. Thus, although the dissertation is mainly a contribution to moral philosophy it also sheds some light on topics in epistemology.

The dissertation defends the thesis that knowing oneself first-personally, as opposed to third-personally, is essential for developing virtue. In arguing for this thesis I am attempting to do justice to an insight recognized by most forms of psychotherapy but which has been neglected within virtue ethics (as well as within social psychology and cognitive science). The work of David Finkelstein and Richard Moran provides the framework for an argument I make against a number of social psychologists and cognitive scientists working on the topic of self-knowledge. I argue that if the aspiration to know oneself is part of a person's ethical development, then it cannot be merely an aspiration to acquire information about oneself but also, and quite importantly, to relate and engage with this information in a first-personal way. I claim that it is only with this kind of self-knowledge that one can aspire to unify rationally our competing and conflicting perspectives about how to live one's life into a coherent whole.

I also appeal to the work by Jonathan Lear to elucidate further the nature of first-personal self-knowledge. Moran and Lear agree on the thesis that a certain kind of first-personal self-knowledge is important for ethical development but disagree on how to characterize this kind of self-knowledge. Lear defends an expressive model of first-personal self-knowledge according to which a person has first-personal self-knowledge of her mental state in virtue of the fact that she can *express* it in a self-ascription. Moran defends an agential model of first-personal self-knowledge according to which a person has first-personal self-knowledge of her mental state if she can make up her mind by judging whether it is merited to hold such a mental state. Although their positions seem in tension with each other I argue that they are actually complementary. Each of these types of first-personal self-knowledge plays a distinctive role in our ethical development. Merely-expressive self-knowledge is required to make explicit patterns of thought that interfere with a person's ability to live well. Agential self-knowledge is the ultimate aim of the kind of self-transformation. A properly human life is one in which a person's mental states are formed and can be transformed though her rational

assessments of whether it is merited to holding them. These two ways of knowing oneself are connected by the fact that, if things are working well, the judgment about these merits of holding these mental states leads to the actual having these mental states. In fact, I argue that the expressive dimension of self-knowledge facilitates the process whereby instinctive and automatic responses that are not rational become unified with, and transformed by, our self-conscious reflection. This process ultimately leads to the acquisition of the kind of agential self-knowledge that Moran defends.

Examining these varieties of self-knowledge of our mental states help us to come to a better understanding of the way in which self-examination and self-knowledge contribute to the development of virtue. It also brings to light the ethical significance of a distinction that is important in epistemology and, in doing so, also sheds some light on discussions of self-knowledge within epistemology. Finally, by discussing the contributions to the topic of self-knowledge by psychotherapy, social psychology and cognitive science, this examination allows me to offer a textured account of ethical development that is empirically informed by the contributions to the topic in the 20^{th}century, contributions that are necessary to provide a proper account of the development of virtue but which are often overlooked by virtue ethicists influenced by traditional historical figures.

I will start the introduction to the dissertation by establishing a series of distinctions and laying out some key presuppositions that will allow me to sharpen the dissertation's topic. I will then provide a literature review on the topic and will conclude by laying out a sketch of the particular aims in each of the dissertation's chapters.

I.1 Sharpening the Topic

I.1.1 A Thin Conception of Virtue

Because I am trying to make my account as ecumenical as possible, I will refrain from advancing any substantive account of the nature of virtue. I will work, instead, with a thin conception of virtue according to which "virtue" is merely a placeholder for "human excellence." My account is thin because it leaves open how 'excellence' ought to be cashed out. As a consequence, the views I put forth should speak to traditional accounts of virtue which identify virtue with morality (like Aristotle's,[1] Hume's[2] or Kant's,[3]) but also to amoral or immoral conceptions of virtue) like Callicles'[4] or Nietzsche's[5]).

It might be worth emphasizing, however, that leaving open the conception of virtue within which I want to operate does not imply that virtue, as I conceive it, is not constrained by normative standards. The life of the virtuous person can be seen as an embodied answer to the question "How should I live?" And this answer can be mistaken. Even if one puts forth subjective conceptions of virtue, it will be important for my account that such conceptions allow for the fact that the person can, in principle, be wrong about what she takes to be the life that is worth living. This will be important because the fact that a person can be mistaken about the normative standards according to which she ought to live entails that these standards are responsive to reasons.[6] Thus, the model of virtue with which I will work in the dissertation is not one according to which virtue is a mere aggregate of arrational

1. Aristotle 2002.

2. Hume 1998, 2007.

3. Kant 1996.

4. Plato 1987.

5. Nietzsche 1974.

6. With this I am not attempting to put forth the view that such answer should be reached deductively. What I am proposing is that such an answer ought to have rational coherence and can be defended rationally.

sentiments and attitudes. In this sense my account is incompatible with certain sensibilist readings of Hume's and Nietzsche's.

Cashing out virtue in terms of excellence and not in terms of morality should bring to mind the distinction, initially drawn by Bernard Williams, between ethics and morality, a distinction which has now become almost standard in anglophone moral philosophy.[7] The main question with which the dissertation's protagonist is concerned is not the question "How should I be morally good?" but "How should I live an excellent human life?," where the answers to these two questions may come apart. Thus, the dissertation is mainly about ethics and not about morality. And although it has been standard for moral philosophers to argue (or assume) that such a life is a moral one, in the dissertation I will not be making such an assumption.

I.1.2 The Dissertation's Hero

The main character of the dissertation, the dissertation's protagonist, is a person who is not yet virtuous but who is trying to become virtuous. I will often refer to this character as the "ethical pilgrim." This locution pays homage to Iris Murdoch, one of the first 20^{th} century philosophers to offer what we now consider to be a virtue ethics approach to moral philosophy. She was also one of the first philosophers in the last few decades to pay scholarly attention to moral development and moral education. She coined the expression "moral pilgrim" to refer to the person who is aspiring and struggling to become moral. I find this expression to be quite compelling since it brings to the fore that this struggle for self-improvement is a journey or quest of special significance. Because my dissertation will focus on ethics and not morality, I will refer to this character, the dissertation's hero, not as the "moral pilgrim" but as the "ethical pilgrim." On occasions, however, I will also refer to her as the "ethical aspirant", especially when I want to highlight the fact that she is someone aspiring to become

7. Williams 2000, 1985.

ethical.[8]

I.1.3 A Few Distinctions

The main claim of the dissertation is that the ethical pilgrim should come to know herself better, in particular, she should come to know those mental states that are interfering with her capacity to be virtuous. There are different ways in which one can hear this claim. Distinguishing these will be important to get a proper grasp of the aims of the dissertation.

Self-Examination and Self-Knowledge

The first thing that might need to be clarified is the relationship between self-examination and self-knowledge. The aim of self-examination, at least of the kind of self-examination in which we will be interested, is self-knowledge. Thus, whenever I use "self-examination" I will understand it as a form of examination that *aims* to bring about self-knowledge. As a consequence throughout the dissertation I will be assuming that self-examination is not valuable in and of itself but that it gets its worth because it brings about self-knowledge.

8. It is worth mentioning that even though the idea of an ethical aspirant is familiar to all of us, there are significant difficulties in properly conceptualizing this idea. Donald Davidson, in fact, argued that aspiration was a form of irrationality (Davidson 1982). The difficulty, as Agnes Callard has suggested, is that before the ethical pilgrim has become virtuous, she is trying to see the world in the way a virtuous person would. Arguably, however, if the aspirant seeks to see things in the way that a virtuous person would, she must already see that that is the way to see things. And doing so seems to presuppose that she can actually see them as the virtuous person would, it would presuppose that she is virtuous (Callard 2017 (Forthcoming), Chapter 3). The puzzle, in a more abstract way, can be formulated through the following question: how can a person ascertain the propriety of a point of view which she does not occupy? It is beyond the limits of the dissertation to address these difficulties. (Callard 2017 (Forthcoming), however, has a compelling solution to it).

Generic and Concrete Self-Knowledge

The injunction "know thyself!" will be heard, by a typical contemporary western reader, as an injunction to come to know the intricacies of her particular, individual, psyche. This way of understanding this injunction has now become natural for those of us who have been raised in a post-Freudian society, a society where psychotherapy is ubiquitous. The injunction "know thyself," however, can also be heard on a generic register. It can be taken to be an injunction to come to know myself generically, to know the kind of creature that I am.

And it is worth noting that, as a number of scholars have convincingly argued, during a good deal of our history such an injunction was usually heard in the second way.[9] They have demonstrated that the most famous injunction on the topic in the West, the Delphic injunction "Know thyself!," was intended in this generic register. It was meant to raise the question "Who am I *qua* human being?" The answer to this question was meant to bring about the recognition of our human limitations *vis-à-vis* the power of the gods. It was meant to work as an invitation for the person to recognize what it is to be a human being in general and not as an injunction to come to know the particularities of her individual soul.

But even even if ancient writers usually conceived of "self-examination" and "self-knowledge" in a generic register, it is not adequate to think that they only understood these injunctions in such a register. Some of the scholars who have shown that ancient references to self-examination and self-knowledge operate in the generic register have been so impressed by the pervasiveness of the generic dimension of self-knowledge in antiquity that they have suggested that the individual forms of self-knowledge which are pervasive in our culture are actually a modern invention. Michel Foucault has often been read as arguing for this.[10] His

9. See, for instance, Foucault 2005; Annas 1985; Johnson 1999.

10. Foucault 2005.

recognition of the pervasiveness of the generic aspect of self-knowledge led him to say that concepts like "subjectivity," "individuality," "personality" and "self" are modern creations, absent in the philosophical landscape of antiquity. A good place to look to see how this takes things too far is Plato's Alcibiades.[11] Benjamin Rider has shown that, even though there are important strands in the dialogue where self-knowledge is meant to be understood in what I am calling a generic register, there are also very significant strands where it needs to be understood in the register of the particular individual.[12] As such, he shows that the individual register of the question was available and operative in antiquity.

This should not be surprising. The generic and the individual registers are interconnected. Even if people in different historical periods have been prone to hear the injunction in a generic or an individual register, these two registers are always intertwined and interact. The fact that these two aspects of self-knowledge are intertwined has been emphasized by a number of virtue ethicists. These scholars have actually highlighted the back and forth interactions between these two forms of self-knowledge.[13] Our (perhaps tacit) generic understanding of what it is to be a flourishing human being guides our efforts towards self-improvement. But our improvements do not leave these ideals untouched; they often reshape them. As our own capacity to be virtuous increases, our conception of the kind of person that we are meant to be also shifts.

Thus, it is not appropriate to establish a sharp contrast between an ancient impersonal or generic conception of self-knowledge and a modern individualistic conception. Even if different historical periods have placed emphasis on the former or the latter, these two forms of self-knowledge are not two alternatives from which one is supposed to choose, but complementary aspects of one single endeavor. Many contemporary discussions about self-

11. The question of the dialogue's authenticity is besides the point here. What matters here is that this is an ancient text that reflects an ancient Platonic outlook.

12. Rider 2011.

13. See, for instance, Murdoch 1997a; Annas 2011N. Snow 2016.

knowledge fail to be aware of this.[14]

Self-Knowledge of the Ideal and of the Actual

A further distinction that is important to establish is the distinction between knowing what I will call, for lack of a better term, the "ideal self" and the "actual self." The ethical pilgrim, unlike a brute, asks herself how her life ought to be lived. Her response provides an ideal according to which she is meant to live and to which she is trying to conform. I call this ideal her "ideal self". The ideal self is displayed in the person's aspirations, aspirations which are often at odd with her actual behavior. The "actual self," is not displayed in the person's aspirations, but in her current behavior, behavior which is sometimes at odds with her aspirations.[15]

Before continuing allow me to explain how the distinction between ideal and actual self bear on the previous distinction between generic self-knowledge and individual self-knowledge. Generic self-knowledge is a part, but a proper part, of the ideal self. Generic self-knowledge is about the standards of what human beings should be like.[16] As such,

14. One often comes across references to the Delphic injunction or to Socrates' concern for self-knowledge in accounts that describe concrete individual self-knowledge. These anachronic references can be found in disciplines as different as philosophy (Deweese-Boyd 2012; gertler 2011), psychology (Timothy D. Wilson 2002; Timonthy D. Wilson and Vazire 2012, neuroscience (Lieberman 2012) and psychotherapy (Falkenström 2012). These discussions are referencing the oracle or Socrates perhaps as a way to show the venerable tradition within which their account is positioned. But what this reference shows is an anachronism that betrays a lack of awareness about the topic that they are discussing, a failure to see that the kind of self-knowledge about which they speak is not the kind of self-knowledge with which the injunction in the Delphic oracle's was concerned. The problem in failing to note this distinction is not that it is historically inaccurate but that it hinders the writer's and reader's capacity to properly understand the different ways in which self-knowledge can contribute to the development of virtue.

15. Although the distinction I am drawing is extremely familiar, I do want to flag the fact that the way I which I am cashing it out is imprecise. The distinction between aspirations and current behavior is not exhaustive. Some of the person's behavior (say, the person's verbal reports of her aspirations) are actually manifestation of the person's ideal self not of her actual self. Because it will not be important for the dissertation to distinguish between actual self and ideal self with total clarity, I have not attempted to offer a more precise characterization of the distinction.

16. According to some ethical outlooks, the fact that we are human beings does not determine, in any way, how we should behave. If this is the case, then generic self-knowledge would be vacuous in these approaches.

generic self-knowledge reveals what any human being should aspire to become, it is knowledge of part of the ideal self. But generic self-knowledge is only a proper part of the ideal self because the ideal self outstrips the generic ideal. There are aspects of the ideal self that have to do with standards or ideals that are not generic to human beings but which are specific and apply only to an individual person. My aspirations to be a teacher or to be a person with a mellow temperament are not generic aspirations to which every human being should aim. Even though they might be central to my own life project, these aspirations are specific to me in particular and not to me *qua* human being.

The injunction "Know thyself!" can be understood as making a reference to either pole. It can mean an injunction to reflect on and reconsider the type of person that one ought to be (as a human being or as an individual) or an injunction to come to know some features of one's actual self about which one might be ignorant. This suggestion goes against (or at least puts some pressure on) a number of recent influential accounts of human identity such as Harry Frankfurt's, Christine Korsgaard's, and Angela Smith's.[17] These accounts propose that the person should be identified with the mental states that she reflectively endorses, that is to say, the mental states to which the person would like to adhere, the mental states of her ideal self. In the dissertation I will argue against these views by showing that one cannot properly grasp who a person is, the kind of life that she is living, if one does not see her as operating between these two poles. The actual self and the ideal self are two aspects that characterize a person's life. Ignoring either of these gives one a lopsided picture of such a life.

The actual self need not be out of sync with the ideal self. In fact, when things are going well, actual and ideal are meant to coincide or at least to come close to one another. Generous people also believe that they should be generous. For them, the aspiration to be generous and the fact that they are generous are of a piece.[18] The dissertation will be concerned,

17. Frankfurt 1988; Korsgaard 2000, 2009; A. M. Smith 2005.

18. This coincidence corresponds to what Aristotle called *energeia* (Aristotle 1999). It is nicely illustrated

however, mostly with cases when this does not happen, cases where "who I actually am" is misaligned with "who I want to be." It will be mainly concerned with understanding the role that self-knowledge plays in our attempts to conform to our ideal self, with understanding how coming to know those aspects of ourselves which conflict with our ideal self helps us achieve this ideal selves.

I.1.4 A Central Supposition

These distinctions allow me to sharpen the dissertation's aims and topics. As I have said above, the dissertation's protagonist is a person who is not yet virtuous but who is trying to become virtuous. Her struggle to become virtuous involves an interaction between two poles: an ideal aspiration, to which she is attempting to conform, and the reality of who she actually is, a reality which she is trying to transform.

The correctness of one of these poles, the person's ideals, determine the success of her ethical project. This is worth emphasizing. A person's aspirations to become virtuous are only virtuous if her aspirations turn out to be virtuous aspirations. If one's ideals are vicious, conforming to them will actually make one more vicious, not more virtuous. If self-examination and self-knowledge are, as I will argue, conducive to the attainment of one's ideals, they will contribute to making the person more vicious if her ideals are vicious.

As I noted, the question about what makes one's ideals virtuous is beyond the confines of this project. Describing these ideals amounts to providing a substantive account of human excellence. It amounts to responding fundamental questions related with generic self-knowledge, such as "What is it to be a human being?" and "What is a good human life?" These questions are at the center of ethics but it is beyond the limits of the dissertation to address them. I will also avoid articulating or fleshing out what an individual person's ideal self should be. My main interest in the dissertation will is to understand the process

with the example of a healthy person: a healthy person is healthy, but being healthy involves an aspiration and various efforts to be, and continue to be, healthy.

whereby, by acquiring self-knowledge, the person transforms her actual self in the direction of her ideal self.

But even if I barely say anything substantive about the ideals that our protagonist ought to pursue, I will focus on cases in which the agent's ideals are mostly correct. It is worth saying that it is somewhat artificial to leave fixed the person's virtuous ideals and to assume that these ideals are reasonably correct. As I mentioned earlier, our ethical growth informs and reshapes our ideals. Moreover, an important part of the process of ethical development involves, precisely, coming to acquire the right ideals. My aim in the dissertation, however, is not to understand how a person can correct or improve her own ethical ideals but to understand how self-knowledge helps a person transform herself so as to conform to the ideals which she holds. Discussing what these ideals need to be and how a person can acquire the right ones will lead us far afield.

I.1.5 The Dissertation's Main Question

The guiding question of the dissertation, then, is: "How does self-knowledge of her unconscious mental states contribute to the ethical growth of someone who is striving to become virtuous, and who has a relatively correct view of how she should live, both at an individual and a generic level?"

I.2 Literature Review

I.2.1 Moral Philosophy and Virtue Ethics

Many dissertations set themselves in an already ongoing debate where there is a wealth of literature from which to draw. This is not the case with my dissertation. Within the last few decades, there has been very little material on this topic in the anglophone tradition.

There is even less literature addressing the role that different forms of self-examination and self-knowledge play in the development of virtue.

Virtue ethics has been usually interested on topics related with moral education and ethical development. Nearly all scholars who belong to this tradition agree on the thesis that self-examination, and the self-knowledge that it brings about, is important for the development of virtue. As I mentioned in the opening paragraph, it is perhaps because this thesis seems obvious that contemporary virtue ethicists have not investigated it in any depth. Very little attention has been paid to reflecting, problematizing, and understanding the role that self-examination and self-knowledge play in such development.

There are, however, a few books that survey how the concept of self-knowledge has developed in the Western philosophical tradition. In *Morality and Self-Deception*, for instance, Mike Martin investigates the relationship between self-knowledge and morality through an examination of different approaches taken in the west to the relationship between self-deception and morality.[19] One of the main aims of his book is to shed light on three fundamental values: sincerity, honesty and authenticity. Similarly, in *Self-Knowledge and the Self*, David Jopling explores the alternative answers that different traditions in the history of Western philosophy have given to the question: "What is the nature the self that is the object of self-knowledge?"[20] Finally, In *Giving an Account of Oneself*, Judith Butler attempts to highlight the intimate relationships between self-opacity, responsibility and morality.[21] She surveys the conception of self-opacity in a number of philosophers from the continental tradition, using their views to articulate that how I conceive of myself and what I know of myself, is constituted by my interactions with and against the social world. All of these authors are sympathetic to certain aspects of the project of virtue ethics, although their investigation has not been explicitly pursued within this theoretical framework.

19. Martin 1986.

20. Jopling 2000.

21. Judith Butler 2005.

There are two noteworthy book-length texts in the last forty years which have investigated self-knowledge within the virtue ethical tradition. The first, *Virtue and Self-Knowledge*, written by Jonathan Jacobs more than thirty years ago, argues that having self-knowledge is constitutive of virtue.[22] The argument relies on establishing an internal connection between the capacity for self-knowledge of the person's life narrative and her ability to determine her actions and dispositions through such self-knowledge. He argues that "opacity in one's historical self-understanding undermines self-determination."[23] The second book-length study is a PhD dissertation, *Self-knowledge and Moral Virtue*, written by Kathleen Poorman Dougherty sixteen years ago.[24] In this work, Dougherty focuses on self-knowledge of our character. Within Aristotelian framework she defends that self-knowledge of our character is necessary but not sufficient for being virtuous.

There are two ways in which my approach is different from theirs. First, my dissertation focuses on self-knowledge of our mental states. Jacobs' book focuses on self-knowledge of our personal narrative, that is to say on the person's knowledge of her own life history. Dougherty's, in turn, is concerned with examining the person's knowledge of her own character, not on her knowledge of her own mental states. Second, both of these books focus on the *phronimos*, the person who is already virtuous, and not on the ethical pilgrim who is aspiring and struggling to become virtuous. Thus, while my dissertation focuses on how self-examination and self-knowledge are valuable tools to develop virtue, Jacobs' book and Dougherty's dissertation focus on how self-examination and self-knowledge are constitutive features of the life of the virtuous person.

Most other virtue ethicists who have have dealt with this topic in a sustained way have attempted to criticize the view that self-examination or self-knowledge are necessary for ethical development. Julia Driver and Iris Murdoch are among the best known authors whose

22. Jacobs 1989.

23. Jacobs 1989, p. 4.

24. Dougherty 2000.

work has been used to challenge this view. Driver has argued that there are certain virtues, which she has labeled "the virtues of ignorance," that are at odds with self-knowledge.[25] Virtues like modesty, she argues, require the person to be ignorant about possessing such virtues. To say "I am modest!," according to Driver, is to utter a self-defeating assertion. The modest person is not someone who asserts that he is modest; he is, rather, someone who underestimates his self-worth. Driver argues that if a person knows that he is modest, this is a sign that he is not modest. Her views on this issue have spurred a lively debate within virtue ethics.[26] Although I believe that her challenges have been responded to conclusively,[27] this is not something that I argue within the dissertation. Driver's position, like Dougherty's, focuses on the character of the *phronimos*, not on the mental states of the ethical aspirant. This entails that my claim that self-knowledge is an important tool to help us grow ethically could be compatible with Driver's challenge. One could argue that knowledge of one's own ethical shortcomings is necessary for ethical improvement even if ignorance of one's virtues is necessary for one's ethical excellence.[28] In the specific case of a virtue like honesty, one could argue that to become modest, the ethical pilgrim needs to come to know the mental states which interfere with his modesty, but still hold that when the process of ethical development has been completed, the virtuous person should be ignorant of his own modesty. In fact, one might even argue that it is part of the modest's person's orientation to think that he is not modest, to think that he is a mere aspirant to modesty. It would be, precisely, his attempts to know the mental states that conflict with his modesty that would contribute to his being a modest person.

Iris Murdoch is has offered a strong attack on the view that self-examination and self-knowledge of aspects of ourselves that we ignore is required for ethical development. Her

25. Driver 1989, 1999, 2001.

26. See, for instance, Flanagan 1990; Dougherty 2000; Raterman 2006; Winter 2012.

27. See Flanagan 1990; Dougherty 2000 and especially Winter 2012.

28. I am indebted to Daniel Rodriguez for helping me formulate this clearly.

criticism of self-examination and self-knowledge come up in her philosophical as well as in her literary texts.[29] Murdoch is well known for championing an ethics of vision as an alternative, or at least as a complement, to an ethics of choice.[30] In developing such an ethics Murdoch rehabilitated concepts such as grace and self-transcendence, concepts largely absent in the philosophical moral discourse of her time (and of ours). These family of views are not accidentally connected: Murdoch's aspirations to vindicate the ethical significance of vision are of a piece with her doubts about the power of self-examination to transform ourselves. Murdoch's suspicions about self-examination are connected with her criticisms of views on moral philosophy which place too much emphasis on the will, views which fail to account for passive capacities like perception that are ethically significant. Murdoch's work serves as a reminder that even if self-examination and self-knowledge are indispensable for ethical development, it is important to open up, to be receptive to forms of transformation where we are passive and whose power of transformation are best characterized as caused by grace.

Although Nomy Arpaly does not define herself as a virtue ethicist, her work can be used to put pressure on the view that self-examination and self-knowledge play an important role for the development of virtue.[31] Arpaly challenges the alleged importance and moral significance that philosophers have given to the agent's conscious awareness of what she is doing. Her view can be used to challenge the thesis that I am putting forth. The argument in the dissertation starts, precisely, by responding to this challenge in chapter II. I will explain a bit more fully what this challenge is and how I respond to it shortly (I.3).

I.2.2 Clinical Experience and Empirical Results

In the dissertation I will be drawing on insights provided by the clinical experience of psychotherapists and on the social psychologists' and cognitive scientists' empirical results. I

29. Murdoch et al. 1963; Murdoch 1993, 1997a, 2000, 2001.

30. Antonaccio 2012.

31. Arpaly 2000, 2003, 2015.

draw on this literature for two reasons. First, engaging with this literature allows me to offer a textured account of ethical development that, being responsive to the insights offered by these disciplines, is properly informed empirically and can contribute to the accounts proposed by virtue ethicists inspired by historical figures writing before the 20$^{\text{th}}$century. Second, given the scant literature on this topic within philosophy, this literature provides me with scholarly interlocutors in connection to whose work I can frame my own view. Allow me to elaborate each of these points.

Therapists of a variety of orientations insist that therapeutic improvement requires acquiring specific forms of self-knowledge. As Sigmund Freud wrote: "Knowledge is not always the same as knowledge: there are different sorts of knowledge, which are far from equivalent psychologically."[32] Moral philosophers writing about ethical development rarely mention (and almost never investigate) that the aspiration to know our mental states, if it is guided by ethical considerations, is not merely an aspiration to acquire information about these mental states. It also an aspiration to know them first-personally and, thereby, to relate to these states in a manner that makes it possible to engage and transform them in a distinctive way.

Contemporary moral philosophers have neglected the ethical significance of this distinction. In the occasional (and often passing) remarks that they make on the topic of self-knowledge, they almost always fail to emphasize (and in some cases they clearly overlook) that we are not meant to relate with our self-knowledge in a merely theoretical way. The received view tends to be that self-knowledge allows us to acquire certain information about ourselves that helps us to improve ethically. Take the following example, described by Jeanine Grenberg in a book which is insightful on the topic of self-knowledge. The example comes from a scene in Fyodor Dostoevsky's *Brothers Karamazov* where Katerina is struggling to figure out whether to visit her former lover, Dmitry, in prison. Grenberg writes: "[Katerina]

32. Freud 1963, p. 281.

looked within herself, at the feelings she experienced in her conversations with Alyosha, and at her pattern of inactivity thus far, and put two and two together. [... Katerina concludes that] shame was informing her inactivity. (...) Katerina recognizes the very live possibility that her shame has informed her inactivity, and thus takes that new self-knowledge as further information in changing the way she acts."[33] Grenberg is conceptualizing Katerina's acquisition of self-knowledge as a theoretical achievement whereby self-knowledge provides her with new information, information that allows her to make a practical decision. This account might be well and good. But I am interested in a dimension of self-knowledge that does not come into view in this kind of remark. Katerina likely relates to her self-knowledge in deeper ways than those described by Grenberg. If things are working well, Katerina's recognition of her pride will not just provide her with information that can now become part of the content of her deliberation. Recognizing her pride might allow her to engage with such pride in such a way that she can engage its warrants in a deliberative spirit, thereby contributing to transform it.

Also within social psychology it has become commonplace to think of self-knowledge as a mere theoretical achievement. Research in this area has shown significant effects in our behavior caused by what seem to be minor and irrelevant details from our environment. Philosophers inspired by these results have been drawn to a picture where the aim of self-knowledge is to identify all of the potential situational factors that can influence our judgment and behavior and to shape our environment so that we can control the way in which these influence us.[34] Very few scholars, to my knowledge, have paid attention to the significance of thinking of self-knowledge as something more than a mere theoretical achievement.[35] The dissertation aims to remedy this neglect.

33. Grenberg 2005, p. 237.

34. See, for instance, Merritt 2000; J. Doris 2002; Samuels and Casebeer 2005; Sarkissian 2010.

35. The few exceptions that I know of are: Nussbaum 1990a; Lear 1990, 1998, 2003; Lear et al. 2011; Flanagan 1991; Richard Moran 2001, 2011; Vice 2006; McGeer 2007.

Philosophers influenced by psychotherapy (most of whom actually draw, specifically, on psychoanalysis) are among the few scholars that seem to be sensitive to the fact that this distinction is important for ethical development.[36] However only two of these authors, Jonathan Lear and Richard Moran, articulate in any depth why certain forms of self-knowledge are ethically superior to others. Lear and Moran offer different, and allegedly competing, accounts of first-personal self-knowledge and of how it contributes to ethical development. I engage with their accounts in chapter V. My suggestion, novel within the field, is that there is more than one phenomena that one can call first-personal self-knowledge, there are at least two different ways to know our minds from the inside, endorsed first-personal self-knowledge and merely-expressive first-personal self-knowledge. I argue that ethical development requires that we cultivate both of these capacities to know ourselves first-personally as each of them play a distinctive role in the process of ethical development.

As I mentioned above, very little work has been done within philosophy to address the main questions of the dissertation. This topic, however, has popularized in psychology in the last 20 years.[37] The seminal article in this literature is, of course, Nisbett and Timothy D. Wilson (1977). Twenty five years after the publication of this article one of its coauthors, Timothy Wilson, published *Strangers to Ourselves*.[38] In this book he sharpened and developed the ideas put forth in the original article. A wealth of research has followed and there are now a variety of research programs in the area.[39] Within psychology, research

36. See, for instance, Martin 2006; Wollheim 1984; Nussbaum 1990b, 1994, 2001; Lear 1990, 1998, 2000, 2003, 2006; Lear et al. 2011; Sherman 1991, 1995; Cottingham 1998; Jopling 2000; Richard Moran 2001, 2011; Jopling 2008; Lacewing 2008, 2014.

37. One of the first noteworthy books within psychology on the topic, published almost fifteen years ago, starts by saying that at the time that the book was published there was very little literature in academic psychology on this topic: "There are few college courses on self-knowledge and few books devoted to the topic, if we rule out self-help books and ones devoted to the psychoanalytic point of view." (Timothy D. Wilson 2002, p. vii).

38. Timothy D. Wilson 2002.

39. A sample of some of these have been collected in the *Handbook of Self-knowledge* (Timonthy D. Wilson and Vazire 2012).

on self-knowledge has developed hand in hand with research on associative learning, priming influences and dual-process theories of cognition, research that shows the importance and pervasiveness of unconscious mental processes in our mental and practical lives.[40] I draw on this research both as a source of insight but also as a foil. In chapter II I rely on this research to make my argument against Arpaly's challenge. I also use some of its results in chapter IV, where I try to bring out an important and unargued premise in the work of epistemologists who defend an agential conception of self-knowledge. In chapters III and IV, I also rely on this research as foil against which to frame my positive view.

Even though moral philosophers, particularly virtue ethicists, have recently become interested in moral education, their focus has often been on the process whereby children grow into moral adults. Little attention has been given to the kind of moral education with which the dissertation is concerned, that is to say, with what one could call "adult ethical re-education." This kind of endeavor, however, is at the center of therapeutic approaches to mental health. Not surprisingly, scholars interested in adult ethical re-education have looked into psychotherapy and drawn insights from it.[41]

I.2.3 Epistemology

Nearly all of the attention that contemporary anglophone philosophers have devoted to the topic of self-knowledge has taken place within philosophy of mind and epistemology. Their attention, however, has focused on explaining the alleged immediacy, authority and privileged access that subjects have to their own minds. In articulating this allegedly privileged form of self-knowledge, they have often focused on narrow aspects of self-knowledge that seldom have any ethical significance: knowledge of our own pains, our sensations or our beliefs

40. For a sample of this research, see Hassin, Uleman, and Bargh 2005.

41. See, for instance, Martin 1986, 2006; Wollheim 1984, 1991, 2003; Nussbaum 1990b, 1994, 2001; Lear 1990, 1998, 2000, 2003, 2006; Lear et al. 2011; Sherman 1991, 1995, 2007; Cottingham 1998; Jopling 2000; Richard Moran 2001, 2011; Jopling 2008; Lacewing 2008, 2013, 2014.

about simple things like whether it is raining. These mental states are seldom of ethical significance. It is, therefore, not surprising that the debates that attempt to account for our alleged first-person authority have been mute about the ethical dimension of the distinction between first- and third-personal self-knowledge. Some epistemologists have gone as far as asserting that these distinctions are, actually, not ethically significant.[42] On the whole, only a handful of philosophers of mind have paid any attention to the ethical dimension of self-knowledge.[43]

I share with an growing number of scholars the sense that these discussions leave to the side extremely important topics that philosophers should be discussing. In showing that the distinction between first-personal self-knowledge and third-personal self-knowledge is ethically significant, the dissertation puts in contact the work that epistemologists have done on self-knowledge and first-person authority with the work that ethicists have done in moral education and ethical development. Thus, although the dissertation is mainly a contribution to moral philosophy it also sheds some light on epistemology.

In showing that this distinction is ethically significant the dissertation puts in contact the work that epistemologists have done on self-knowledge and first-person authority with the work that ethicists have done in moral education and ethical development. Thus, although the main aim of the dissertation is to contribute to moral philosophy, the process of doing so ends up shedding light on some of these debates within epistemology.

It is important to note, however, that the dissertation is not seeking to offer an account of the metaphysics of the mind that explains how subjects have this supposedly peculiar access to their mental states. My aim is not to provide an account of the metaphysics of first- and third-personal self-knowledge. I am interested in the practical distinction between, on one hand, a way of knowing a mental state that *is taken to be* immediate, private and,

42. See, for instance, Carruthers 2011, p. xi.

43. See, Richard Moran 2001; McGeer 2007; Cassam 2014.

authoritative, despite lacking the kind of grounds that we demand of any third person who wishes to ascribe a mental state to another person, and on the other, a way of knowing a mental state that is justified in grounds to which any third person could appeal to, at least potentially. But although I am not concerned with characterizing the ontological status of self-knowledge that is known with alleged first-person authority, my account does establish some constraints that any theory of the metaphysics of self-knowledge would need to respect.

I.3 An Outline of the Chapters

Chapter 2: 'Ignorant Virtue' and 'Knowing Virtue'

Chapter II establishes that self-examination and self-knowledge of our mental states is necessary for the development of virtue. It does so by responding to an objection inspired by the work of moral philosophers such as Iris Murdoch and Nomy Arpaly. According to this objection, many good people develop ethically and act well despite not being very good at examining or knowing themselves.

I start the chapter by showing that even though such people have often been conceptualized as possessing natural virtue, the objection is stronger if we think of them as people with what Arpaly calls 'ignorant virtue.' I then argue that even though the person with ignorant virtue might be more virtuous that other kinds of people, like vicious people or people with what Arpaly calls misguided virtue, the fact that this person is incapable of properly putting his mental states in words and to articulate how they contribute to a life well lived entails that he is not fully virtuous. To show this I rely on research within empirical psychology that establishes that there are a number of tools and opportunities for ethical development that are available to the person with knowing virtue but not to the person with ignorant virtue, tools that are essential for such ethical.

To establish this last claim, I disambiguate between different senses of conscious and

unconscious mental states and clarify the particular type of mental states that are important for my argument, namely, mental states that are meant to be responsive to our own judgments about the merits of having them.

With these distinctions under our belts I proceed bring to bear some empirical evidence to show that ignorant virtue cannot be full virtue. I show that because the motives of the person with ignorant virtue will not figure in his conscious deliberations, these deliberations will be less rational. Some of these motives will also be unavailable in crafting long-term plans. He will be unable to consciously imagine himself fulfilling his virtuous aspirations, and will not be able to use these aspirations to organize and unify mental conflicts that might arise or which manifest only unconsciously.

Ethical development is guided by general principles that often give primacy to long-term goals over the present satisfaction of lower desires or impulses, and we have empirical evidence to suggest that the person with ignorant virtue will often end up privileging the latter over the former. Moreover, being disunified about what he believes consciously and unconsciously will inevitably lead him to sabotage himself; his actions will frequently be faithful to motives that he consciously acknowledges but which undermine his unconscious motives (and *vice-versa*). These conflicts will take a toll on his mental health and well-being, something that will be worsened by his incapacity to talk or write about them. The person with ignorant virtue will not be as capable of integrating his conflicting mental states within a perspective that can envisage his life as a whole and not merely in the here and now. And because the person with ignorant virtue will be unable to formulate her own mental states in words, her unconscious virtuous mental states are likely to lack the proper logical complexity that characterizes genuinely virtuous mental states.

Lastly, the person with ignorant virtue will be unable to properly measure her actions in light of her ideals and she will, therefore, not be a properly self-directed agent. In addition, she will also be unable to have discussions about these mental states with others, depriving

herself of the ability to be transformed by these discussions and to transform others through them. There is reason to believe that being virtuous involves being able to take parts in this.

Chapter 3: First/Third-Personal Self-Examination

Wilson's seminal book, *Strangers to Ourselves*, spurred interest within psychology in the topic of self-knowledge and set the tone for the perspective that psychologists usually take on this topic. One of the key claims defended by him, and which many social psychologists endorse, is that introspection (or, as I will call it, "first-personal self-examination") is extremely unreliable. As a consequence, he suggests that if we seek to know ourselves, we should examine ourselves "from the outside" (or, as I call it, with "third-personal self-examination").

Chapter III is devoted to investigate this view. This investigation will allow us to get clear on the ways in which we can come to know ourselves, to see how first- and third-personal self-examination contribute to the development of self-knowledge and to highlight some of the virtues and vices in each of these strategies to examine ourselves.

A central part of my engagement with Wilson's position (and with this literature at large) consists in developing precise definitions that bring conceptual clarity to the topic. In this chapter I disambiguate between different senses of "introspection" that are often conflated and between different ways to understand the "causes" and "reasons" of our behavior. Getting clear on this is necessary to properly assess the relative merits of first-personal self-examination *vis-à-vis* third-personal self-examination.

The distinctions I draw allow me to explore and analyze Wilson's position. I argue that Wilson's case against first-personal self-examination is significantly weaker than he takes it to be. Neither his theoretical considerations nor the empirical evidence on which he relies entitle him to make his case. What is more, I show that Wilson inadvertently recognizes that there are occasions where it is precisely the cultivation and strengthening of our capacity for

first-personal self-examination that improves our ability to know ourselves.

I conclude the chapter by bringing out a kernel of truth in Wilson's view. There are, in fact, strategies and protocols that can make self-examination more reliable, strategies available only when we examine ourselves third-personally. I argue that these third-personal strategies are particularly appropriate when what is interfering with our self-examination are issues related with our self-worth and our self-image. Reaching this conclusion entails, *contra* Wilson, that what sometimes makes third-personal self-examination more reliable than first-personal self-examination has to do with the kinds of motivated irrationality with which self-deception has been traditionally associated rather than with the mechanisms that, according to Wilson, explain the unreliability of first-personal self-examination.

Chapter 4: Endorsed First-Personal Self-Knowledge

A central claim of the dissertation is that not all forms of self-knowledge are equally important for the development of virtue. The person who aspires to be virtuous is not merely seeking to know certain facts about herself. She seeks to know her mental states with what I call first-personal self-knowledge, a form of knowing that allows her to engage and relate with them in a distinctive way that is important for ethical development.

I start the chapter defining first- and third-personal self-knowledge. I show that two ways to examine ourselves discussed in chapter III correspond to the two ways of knowing ourselves. First-personal self-examination, if successful, leads to first-personal self-knowledge and third-personal self-examination, in turn, usually brings about third-personal self-knowledge. I note, however, an important asymmetry: while first-personal self-examination does not lead, except accidentally, to third-personal self-knowledge, third-personal self-examination might (and sometimes should) lead to first-personal self-knowledge.

I then articulate a further distinction between two kinds of first-personal self-knowledge. The first, which I label "endorsed first-personal self-knowledge" requires the person to be able

to express her mental state in a self-ascription *and* to express her endorsement of such state. By contrast, the other kind of self-knowledge, merely-expressive self-knowledge requires the person to be able to express her mental state in a self-ascription but to be incapable of expressing her endorsement of such state.

The following tree lays out how these different types of self-knowledge hang together.

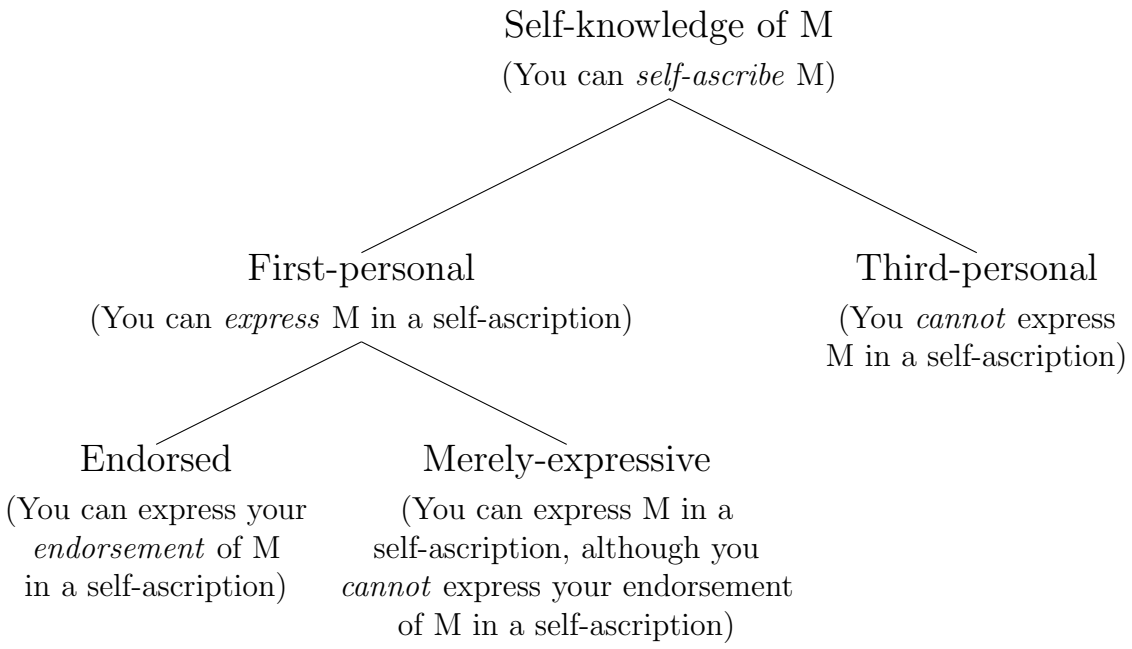

Figure I.1: A Taxonomy of Self-Knowledge

I spend the rest of chapter 4 elucidating why the ethical pilgrim should aspire to have endorsed self-knowledge of her unconscious mental states. I do so by responding to the view standardly espoused by social psychologists according to which all forms of self-knowledge can or should be reduced to third-personal self-knowledge. My argument relies on the work of epistemologists, like Richard Moran or Victoria McGeer, who propose an agential theory of self-knowledge. I argue that endorsed self-knowledge is ineliminable and pervasive. It is not possible to do away with it and, consequently, it is not possible to reduce it to third-personal self-knowledge.

I then articulate why possessing endorsed self-knowledge contributes to the development

of virtue. I argue that to be able to formulate a mental state in speech does not merely involve being able to report it. It also involves being able to defend it when its warrants are criticized and to change it when its warrants are undermined. The person with third-personal self-knowledge can only report a mental state but cannot discuss it with others in a way that opens this mental state to revision; her inability to do this shows the limitations of third-personal self-knowledge.

I also appeal to empirical evidence to argue that our mental states tend to be more rational when we are able to report them with endorsed self-knowledge. Lacking endorsed self-knowledge of our mental states will tend to leave them unhinged from our long-term ethical principles. When you don't have endorsed self-knowledge of your mental states these mental states will tend to be responsive to more immediate desires or to baser inclinations. Finally, acquiring the capacity to report these mental states with endorsed self-knowledge involves a transformation of these mental states from inchoate intuitions that guide the person, either unconsciously or unreflectively, into fully-fledged responses that are embedded in properly reason-giving relationships that fit within her aspirations. Thus, aspiring to achieve only third-personal self-knowledge, relying on decisions based only on third-personal self-knowledge, undermines your capacity to reason consciously, a capacity that is essential to become better at living a flourishing life.

Chapter 5: Merely-Expressive First-Personal Self-Knowledge

In chapter IV I showed how many of the insights that Moran puts forth in his seminal book, *Authority and Estrangement*, can be retooled and applied to my particular project. Moran believes that the insights that he puts forth in the book are supposed to do justice to Freud's practice. Jonathan Lear, a philosopher and practicing psychoanalyst, defends an alternative account that is also meant to be faithful to Freud's practice. Lear criticizes both Moran's account of first-personal self-knowledge and his characterization of psychotherapy. Moran's

account of account of first-personal self-knowledge lines up with what I call endorsed self-knowledge and Lear's with merely-expressive self-knowledge. Examining Moran's position *vis-à-vis* Lear's enables me to analyze the strengths and weaknesses of each of these types of self-knowledge and to assess the place that each of them play in ethical development.

Lear's account helps us to see that the mental states that the analysand is meant to come to know first personally are often recalcitrant to reason. Merely-expressive self-knowledge contributes to make explicit patterns of thought that interfere with the person's ability to live well, facilitating the process whereby instinctive and automatic responses that are not rational become unified with, and transformed by, our self-conscious reflection. Moran fails to see that these recalcitrant mental states, which we cannot know with endorsed self-knowledge, are at the heart of ethical development.

This failure leads him to mischaracterize what takes place within therapy. But while Moran fails to see the instrumental role that merely-expressive self-knowledge plays in therapy, Lear appears to be blind to the normative role of endorsed self-knowledge. Endorsed self-knowledge reveals a unity within the subject of the capacity to speak, reason and act. As such, it sets up the norm of a central feature of a healthy mind. Lear seems not to recognize that merely-expressive self-knowledge is defective and, as such, cannot be the ultimate aim of the ethical pilgrim's aspiration to know her mental states.

The therapeutic aims of Moran and Lear are inseparable from their respective conceptions of the nature of first-personal self-knowledge. In chapter V I contend that, with respect to ethical development, the accounts of Lear and Moran should not be seen as competing but as complementing each other. Each of them help us to understand different aspects of the process of ethical improvement that takes place in psychoanalysis. Endorsed self-knowledge sets up the ideal that should be aimed at in therapy, while merely-expressive self-knowledge provides an important mean that helps the person attain such ideal.

II

'Ignorant Virtue' and 'Knowing Virtue'

It might seem a platitude to say that self-examination, and the self-knowledge that it brings about, is important for ethical development. Recent research in social psychology and cognitive science, however, can be used to question the plausibility of this idea: our capacity to examine ourselves has been shown to be biased and highly unreliable; our unconscious processes have been said to be not only incredibly powerful, but often smarter than our self-conscious reflection; the sheer amount of these unconscious processes appears to make the project of knowing oneself not only daunting but unfeasible; and the possession of self-knowledge has been said to increase depression, thereby interfering with our flourishing. Even within virtue ethics there has been a growing distrust on the supposed importance of self-examination and self-knowledge for ethical development. Scholars in the field have argued that self-examination, which is self-centered, interferes with developing virtue, which requires to attending to the outside world not to oneself; that some virtues, such as humility, require the person to be ignorant about certain aspects of herself; that our efforts to know which aspects of ourselves require moral work are part of fantasies which ultimately do not

allow us to live virtuously; and that the emphasis on self-knowledge offers a reified view of virtue that fails to take into account the many good people who develop ethically and act well despite not being very good at examining or knowing themselves.

In this chapter I intend to address this last challenge, namely, that because some good people are not very good at examining or knowing themselves, knowing and examining oneself must not be necessary for becoming or being virtuous. I start with this one because it is the challenge that allows me to best explain why self-examination and self-knowledge are necessary for the development of virtue. The challenge suggests that if we require the ethical aspirant to examine and know herself, we thereby conceive of the process of ethical development as overtly intellectual and fail to accommodate the possibility of unreflective goodness—the possibility of people who are naturally inclined to the good and who do not require self-examination and self-knowledge to become virtuous. Iris Murdoch is well known for inviting moral philosophers to make room for this alternative. She repeatedly warns in her work about the perils of self-examination. She claims that "the unexamined life can be virtuous"[1] and that goodness "is perhaps most convincingly met with in simple people —inarticulate, unselfish mothers of large families."[2] Murdoch mentions that while contemporary philosophers frequently connect consciousness with virtue[3] it is necessary for them to do justice "to both Socrates and the virtuous peasant."[4]

This view has quite some intuitive appeal, particularly when one is thinking of virtue in the traditional way. As Allison Hills puts it, if you think of the most virtuous person that you know you are unlikely to conjure the image of an intellectual with a really firm theoretical grasp of morality. Most likely, you will picture someone who is not outstandingly articulate

1. Murdoch 1997c, p. 299. See, also, Murdoch 1997d, p. 383.

2. Murdoch 1997b, p. 342.

3. Murdoch 1997c, p. 299.

4. Murdoch 1997c, pp. 299–300.

or theoretically sophisticated, but nevertheless is kind, honest, just, and courageous.[5]

Murdoch's virtuous peasant and inarticulate mother have often been described in the literature as having natural virtue. I argue (and this will constitute the first part of the chapter) that the objection is sharper when we think of these people as having not "natural virtue" but what I will refer to as "ignorant virtue." I am in agreement with Murdoch that to live a virtuous life you don't need to be a professional philosopher adept at constructing abstract theories. But in this chapter I contend that any properly virtuous person needs to have a certain level of critical awareness and powers of self-reflection which are inherent in the capacity to know oneself. Thus, I will argue that the person with ignorant virtue cannot be fully virtuous and, consequently, that ignorant virtue is not proper virtue.

I will begin this chapter by sketching the argument that Julia Annas has provided in *Intelligent Virtue* to defend the view that natural virtue does not amount to full virtue. I will then contrast Annas' response with the account put forth by Nomy Arpaly in *Unprincipled Virtue*. I will show that Annas' response fails to address the challenge of ignorant virtue that Arpaly discusses in her book.

The rest of the chapter will consist in responding to Arpaly. My response will draw on empirical work developed by social psychologists, cognitive scientists and psychotherapists. I will argue, *contra* Arpaly, that the person with ignorant virtue is not fully virtuous: the motivations and reasons on which she acts are not properly rational; she is disunified and disconnected from central parts of herself, parts that are necessary to develop and live a flourishing human life; and there are some central capacities in a flourishing human life that are not available to her. To develop virtue and to be virtuous one needs to know one's mental states. And because there are mental states that are important to our lives but obscured to us, mental state that interfere with our capacity to live a flourishing life, knowing them is essential to living well and must be acquired through self-examination. My conclusion,

5. Hills 2015, p. 15.

then, is that self-examination and self-knowledge of our mental states are necessary for the development of virtue.

II.1 Ignorant (and Inadvertent) Virtue

II.1.1 Annas on Natural Virtue

In *Intelligent Virtue* Annas argues that natural virtue is insufficient for full virtue. Examining her view will provide a window into a traditional way in which "natural virtue" has been understood and will provide us with a compelling rejoinder to the view that natural virtue suffices for virtue. Annas claims that, according to Aristotle, a person has natural virtue when she has a natural predisposition to behave in a way that corresponds with virtue, appearing to behave in the way that the virtuous person typically does. In a number of places of her book, Annas characterizes this kind of natural disposition as a 'knack' or routinized disposition.[6]

One of Annas' main criticisms of the view that identifies natural virtue with proper virtue is the following: "because the person [with natural virtue] lacks the ability to demand, and give, reasons for what he does, he is not equipped to deal with new and unforeseen circumstances. Never having learnt not to take people at face value, for example, the 'naturally brave' person may get into a serious fight over a slight intended as a joke, while the 'naturally sympathetic' person may find herself the victim of scams. As Aristotle says, they are like powerful but blind people who stumble and fall over."[7]

Annas grants that we learn a great deal by copying role models and by imitating our teachers, but she mentions that if this is all that we do, we will merely end up copying the role model or the teacher's mannerisms and style in a routinized way, arguing that a "central

6. Annas 2011, pp. 13–18, 20–27, 70–77.

7. Annas 2011, p. 25.

feature of routine is that the reaction to the relevant situation is always the same."[8] She contrasts this with practical skill and virtue, which "require more than predictably similar reaction; they require a response which is appropriate to the situation instead of merely being the same as that produced in response to other situations."[9] Annas argues that it is only when we know why an action should be done that our action can count as being properly intelligent, that it flows out of true understanding and not merely of conventionally accepted ways of acting.[10] In fact, Annas argues that it is this understanding of why something ought to be done, this understanding of the reasons that justify a particular action, that allows us to exercise virtue in a self-directed way.[11] If one is merely copying or imitating an authority the justification of one's actions will be dependent on the justifications and reasons that such authority might provide for her actions. It is the person's proper understanding of her reasons that allows her to be independent and self-directed.[12]

II.1.2 Arpaly on Inadvertent Virtue

Annas' argument has a number of virtues but her criticisms miss a more interesting and attractive character for my purpose, the person of ignorant virtue. Conceptualizing natural virtue as a form of knack and routinized activity does not do justice to the kinds of capacities possessed by human beings who are unable to verbalize such capacities. There are many examples that one can give where one acquires very sophisticated (and intelligent) skills by imitation and training alone, without the need to consciously understand the reasons that justify the activities involved in such skills. We learn to speak a language without understanding the rules that ground its grammar. Many poets were never taught how to

8. Annas 2011, p. 15.

9. Annas 2011, p. 15.

10. Annas 2011, p. 54.

11. Annas 2011, p. 18.

12. Annas 2011, p. 27.

write poetry by an experienced master, would be unable to articulate how to write good poetry, and are incapable of teaching the art to a pupil. They often learned by reading and writing poetry without an explicit attempt to articulate what they were internalizing. And this does not make them worse as poets. Annas seems to be insufficiently sensitive to the power that imitation has to produce intelligent and sophisticated behavior. If the world provides a human being with sufficient and adequate feedback, this person can often learn merely by doing, without the need to reflect on why some of her doings are successful and others are not.[13] This point does not merely hold about skills that are neutral from an ethical point of view. It is also relevant for ways of dealing with the world that are at the heart of ethics. Glen Gabbard and Drew Westen, two clinical therapists, argue that essential aspects for therapeutic improvement in psychoanalytic interventions (which are meant to lead to a more flourishing life) take place unconsciously, without the awareness of the patient.[14] Nancy Snow has made similar remarks about goal-directed behaviors: "Automaticity researchers are clear that nonconsciously activated goal-directed behaviors are not reflex reactions to stimuli, but are intelligent, flexible responses to unfolding situational cues and display many of the same qualities as consciously chosen actions."[15]

In *Unprincipled Virtue* Nomy Arpaly portrays a very attractive set of exemplars of what one might call people with "ignorant virtue." If one wants to understand why self-examination and self-knowledge are central to the development of virtue one will need to be able to show, not just that 'natural virtue,' as described by Annas, is insufficient for virtue but also that 'ignorant virtue' does not amount to proper virtue.

Arpaly's work brings out a very important family of cases frequently ignored by moral philosophers. She remarks, rightly so, that within discussions of moral philosophy, "[a]gents are often described as acting for reasons that they endorse or as blindly following their desires,

13. Baumeister and Bargh 2014, p. 40.

14. Gabbard and Westen 2003.

15. N. E. Snow 2006, p. 548.

as if these were jointly exhaustive categories."[16] Her book is meant to provide a corrective to this by discussing examples where people are neither acting for reasons that they endorse nor blindingly following their desires. A central example in her book, and which will also play a significant role in this chapter, is Huckleberry Finn. According to Arpaly's reading, which I will simply follow here,[17] Huck has decided to turn in his escaped slave friend, Jim, as soon as he has the opportunity. When such an opportunity arises, however, Huck discovers that he cannot do it. Huck believes (or, to be more precise, *consciously* believes) that the right thing to do is to turn in Jim. But despite his conviction he is incapable of following through on his decision. Arpaly offers a compelling reading according to which Huck's actions were neither the works of mere inclination nor a mere lucky accident of temperament. On the contrary, Arpaly suggests that Huck, through his long acquaintance with Jim, gradually comes to know (albeit in an unconscious or non articulate way) that Jim is a full-fledged human being, someone who ought not be treated as a piece of property. It is this awareness, according to Arpaly, that explains Huck's inability to turn Jim in, and why we tend to think of Huck's action as morally praiseworthy.

According to Arpaly's reading of Mark Twain's novel, Huck is acting for a moral motive without knowing that it is a moral motive.[18] Cases like this, she argues, put into question the standard way in which autonomy has been conceptualized as well as the central role that, according to moral philosophers, the first person perspective is supposed to play in our moral life. Although an agent like Huck explicitly endorses his decision to turn in Jim, and although he criticizes himself for his inability to follow through on this decision, this does not entail that the elements of his psychology which prevent him from turning Jim in are not

16. Arpaly 2003, p. 28.

17. It is not really important for the kind of argument that Arpaly is putting forth (and for my own) whether her reading of the novel is correct. The point is that the character she describes is one we can recognize as being real.

18. Arpaly 2003, p. 10.

his.[19] Arpaly helps us to see that, even though the motives that hindered Huck from turning in Jim are unconscious, they nevertheless correspond to central parts of Huck's psychology. Consequently, the actions that follow from them should be attributed to Huck, they should be conceptualized as *his* actions.[20]

I am very sympathetic to Arpaly's project. She does an excellent job of helping us see that it is possible for an agent to be more rational by acting *against* his best judgment.[21] According to Arpaly, the fact that a reason is conscious is neither necessary nor sufficient for it to count as a good reason or to rationally justify an action. As such, Arpaly is proposing a theory of human rationality that takes into account not only conscious, but also unconscious mental states.[22] And in doing this, she takes herself to be rectifying a tendency within discussions in moral philosophy where "the unconscious does not usually get its due."[23] I find myself in agreement with a good deal of this.

Arpaly explicitly states that the examples that she is providing are not examples of what is usually called "natural virtue."[24] When she denies that her account should be identified with "natural virtue," she is merely distancing her proposal from accounts like Annas, which conceive of "natural virtue" as the outcome of certain atavistic mechanisms, animal tendencies or routinized knacks.[25] According to Arpaly's account, any virtuous person is able to respond properly to a moral situation, not because of a reliable tendency that happens to bring it about, but because the person is moved by the recognition of the right-making features of the action, by the features that make it morally worthy. It is not an accident, according to Arpaly, that Huck is acting as he does. His actions are also not arising

19. Arpaly 2003, p. 14.

20. Arpaly 2003, p. 6.

21. Arpaly 2003, p. 36.

22. Arpaly 2003, p. 33.

23. Arpaly 2003, p. 29.

24. Arpaly 2003, pp. 9, 76; Arpaly 2007, p. 429

25. Arpaly 2003, pp. 9, 76.

from the mere sympathy of a good seeing-eye dog. According to Arpaly, Huck refuses to turn in Jim because he recognizes that Jim is a human being, deserving to be treated with the dignity of a human and not as a piece of property that can be owned. To the extent that this is so, Huck's resistance to turn in Jim flows from a recognition (albeit unconscious) of the right-making reasons that make his action morally worthy.

She does not spell out in any detail how she understands the term "natural virtue." From the little that she says, it seems that she understands it to be a certain way of being that provides the person with a way to respond to situations that happen to align with virtue in a somewhat accidental way,[26] either because of some atavistic mechanism, or thanks to the reliable tendencies akin to those of a good seeing-eye dog.[27] But her account is an account of what I have been calling "ignorant virtue"

At the heart of Annas' criticism is the view that the person with natural virtue is not guided by reasons but rather by routinized responses that make her unable to respond to "new and unforeseen circumstances."[28] As I mentioned, Annas' seems to be overlooking the impressive capacity that we have to process and assimilate information unconsciously. She seems to overlook the fact that we can respond in highly intelligent and responsive ways that we cannot articulate verbally or know consciously.[29]

Annas suggests that the actions of the person with ignorant virtue flow out of conventionally accepted ways of acting.[30] Arpaly's examples show that this is false. Huck is someone with ignorant virtue whose actions go against the conventional racist responses of his time. Annas also argues that it is only when we know why an action should be done that our action

26. Arpaly 2003, p. 76.

27. Arpaly 2003, p. 9.

28. Annas 2011, p. 25.

29. These unconscious capacities are so impressive that some social psychologists have gone as far as to argue that "everything can be primed, (...) conscious intent and guidance is not necessary for even the highest of higher mental processes to operate" (Bargh 2007, p. 2).

30. Annas 2011, p. 54.

can count as being properly intelligent.[31] Arpaly does not disagree with this. Her examples, however, are meant to show that this knowledge might not be explicit or articulable by the person. According to Arpaly's reading, Huck is not acting blindly; he is acting from reasons that reveal that he has a true moral understanding of the situation at hand.[32]

II.1.3 Arpaly's Challenges to My View

It is possible to use the account put forth by Arpaly to challenge the main thesis that I am putting forth in this chapter, namely, that self-examination and self-knowledge are necessary for ethical development. Arpaly is holding on to the idea that morally worthy actions require the person to recognize the right-making features of these actions. But Arpaly is opening conceptual space to think that this recognition may not be explicit and many not be articulated by the person. As Robert Pippin has highlighted, Arpaly is committing herself to forms of what one might call "unknowing knowingness:" "'Unknowing' because their morally relevant reasons are not consciously invoked or consciously attended to; 'knowing' because, even if the agent avows contrary reasons, she can still be said to be, in the cases that interest Arpaly, acting 'on' or 'for' the relevant moral reasons."[33] Thus, in criticizing the privileged and necessary role assigned to conscious attitudes and beliefs, Arpaly is criticizing the need for us to think that the virtuous person would be able to verbalize these reasons.

In putting forth such a view, Arpaly takes herself to be challenging "the importance of the first person perspective for moral philosophy."[34] Her view puts into question the alleged

31. Annas 2011, p. 54.

32. The view that our reasons (and the values that they display) can be unconscious has also been defended by other moral philosophers. Thomas Scanlon, for instance, has suggested that, to the extent that we are disposed to behave in a particular manner, we give evidence, not merely that we value the state of affairs at which the action is aimed, but that such state of affairs is reason-giving for us (Scanlon 2002). Similarly, Angela Smith suggests that in cases where the agent is displaying unconscious attitudes, he is judging the actions that flow from them as good in some way. According to her, our attitudes, both conscious and unconscious, express our values and, through them, who we are (A. M. Smith 2005, 2008).

33. Pippin 2007, p. 292.

34. Arpaly 2003, p. 17.

importance and moral significance that philosophers have given to the agent's conscious awareness of what she is doing. Formulated in this way, Arpaly's proposal constitutes a direct challenge to the view I am putting forth in this work. When I claim that acquiring self-knowledge is important for ethical development, what I have in mind is not an "unknowing self-knowledge" but *conscious* self-knowledge, one that the agent can articulate verbally. Huck's example is meant to help us see that responsiveness to moral reasons by an agent need not be conscious, explicit, or understood as such by him; an agent can act morally without knowing it (in fact, even in spite of explicitly mischaracterizing and misdescribing his own action as immoral).[35]

If one grants that a person can act morally without explicit, conscious, knowledge of the morally relevant reasons on which the action is grounded, then it is plausible to think that it is possible to transform oneself ethically without consciously knowing why or how one is transforming oneself. Arpaly comes close to acknowledging this when she proposes that significant life changes that are morally important can occur, not merely in the absence of, but sometimes even in spite of, contrary deliberation. She constructs some of her examples in such a way that the reader is led to think that it is the person herself who is transforming her own mind. By the end of the chapter I will have shown what is problematic with this view. Before doing so, I'd like to complement Arpaly's own suggestions by drawing on a body of research in empirical psychology. This research suggests, not merely that it is not necessary to require the virtuous agent to be conscious of what she is doing and why, as Arpaly suggests, but that this requirement can even be pernicious.

35. There are moments in her book where Arpaly seems to cash out this challenge in terms of deliberation. For instance, she writes: "only from the first-person perspective does the first-person perspective appear necessary for acting rationally: from the third-person perspective, there are people who act rationally without conscious deliberation and hence without a first-person experience of distance from their desires" (Arpaly 2003, p. 20). To the extent that she is pegging her whole account on a particular account of deliberation, her account would not challenge mine. I don't think that this is the most attractive reading of her work. A more ambitious reading challenges, not merely the fact that one can act rationally without *conscious deliberation*, but that one can do it without *conscious knowledge*.

II.2 Conscious and Unconscious States of Mind

Timothy Wilson, a leading scholar within social psychology, has argued that "[t]he causal role of conscious thought has been vastly overstated."[36] He has suggested that human beings are much better at detecting complicated patterns unconsciously than consciously: "Numerous studies on covariation detection show that the conscious system is notoriously bad at detecting correlations between two variables (e.g., whether there is a relationship between people's hair color and their personalities). In order to detect such relationships, the correlation has to be very strong, and people must not have a prior theory that misleads them about this correlation."[37] Wilson offers evidence from a number of experiments to show that we are often capable of detecting many of these patterns unconsciously and that to detect these patterns it is better to keep our consciousness out of the way.[38]

Wilson has also suggested that it is often better to make our decisions guided by what our gut feelings tell us and not by careful conscious reflections on our reasons. The latter, he suggests, often lead us to make wrong choices.[39] A version of this idea has been staunchly defended by Ap Dijksterhuis and Loran Nordgren, who argued in a seminal paper that it is not always advantageous to engage in thorough conscious deliberation before choosing: choices in complex matters (such as between getting different houses or different cars) should be left to unconscious thought.[40]

These results, if correct, require us to revise some of our natural preconceptions about the role that consciousness is supposed to play in our lives. The interpretation of these results provide a valuable corrective to philosophical approaches which have give too much weight

36. Timothy D. Wilson 2002, p. 107.

37. Timothy D. Wilson 2002, p. 62. See also, Timothy D. Wilson 2002, pp. 26–7.

38. Timothy D. Wilson 2002, pp. 26-7.

39. Timothy D. Wilson 2002, pp. 32–6, 63, 170–3 See also Gladwell 2005.

40. Dijksterhuis, Bos, et al. 2006, p. 1005.

to our consciousness and, in doing so, have failed to take into account the significance of unconscious activity in our lives. It is a corrective that Arpaly is, precisely, trying to spur with her work.

However, and as I will argue shortly, even if these results are correct, they should not lead us to think, as some leading psychologists have suggested, that consciousness works as a post-hoc phenomenon that merely tracks behavior,[41] or as a compass that only reports on the boat's direction but plays no part on its steering.[42] There is room to concede what these social psychologists say about the limitations of conscious thought and the power of our unconscious thought and still argue that having self-knowledge, that is to say, being conscious of what we are doing and why, is of central importance for our lives. This is, in fact, what I will proceed to argue. I will defend the following: even if the influences of unconscious mental states are pervasive, and even if there is information that is processed better unconsciously, it is still the case that it is only when one has conscious self-knowledge that one can properly live in an ethical way.

Most psychologists grant that conscious states of mind play *some* role in our lives. Their disagreement lies in the importance and pervasiveness that these conscious states of mind play in our mental economy. And although there have been polarizing debates about this within social psychology in the last two decades, we have witnessed a slowly emerging consensus on the matter in the last few years. In 2014, John A. Bargh and Roy F. Baumeister, two well known psychologists that have been on opposite sides of these debates, coauthored a paper where they laid out some of their significant areas of agreement. I will approach their co-authored paper as laying down the consensus towards which contemporary psychologists are heading.

I will complement these responses by drawing on insights from psychotherapy. These two

41. Jeannerod 2006, pp. 36–7.

42. Wegner 2002, p. 318.

traditions nicely complement each other on some of the key aspects which I will stress. Scholars writing about virtue ethics who draw on empirical work usually circumscribe themselves, either to the work of psychotherapists or to the work of empirical psychologists. Bringing these two disciplines together allows for a richer and more comprehensive perspective. It is worth pointing out, however, that doing this is not without its conceptual difficulties. The scholars working within these different disciplines do not always carve out the conceptual space in the same way. When one moves from one to the other there are shifts in the conceptual boundaries. In this chapter, it will be important to have clarity about the particular way in which each of these disciplines conceptualizes the distinction between conscious and unconscious states of mind. It is to this that I will now turn my attention.

II.2.1 What Is Conscious as What Is Reportable

Let me start by describing how I will be conceptualizing, in the dissertation, the distinction between "conscious" and "unconscious" states of mind. I will say that a mental state is "conscious" if the person can (genuinely) self-ascribe it and "unconscious" if she cannot. My definition, then, ties conscious states of mind to our capacity to speak about our mental states, in particular to our capacity to self-ascribe them. To be conscious of a state of mind is to have a certain capacity. This entails that for a state to be conscious, in my sense, it is not necessary that it is effectively self-ascribed in speech here and now. One can be conscious of a state of mind and not exercise the capacity to self-ascribe it. However this capacity for self-ascription is one that has to be available for me here and now; I should be able to deploy it whenever I want.[43]

43. In other words, and to put it in Aristotelian terms, this capacity is a second potentiality and a first actuality.

II.2.2 What Is Conscious as What Is Mentally Occurrent

Within the psychological literature, particularly the one that I will discuss here, a conscious state of mind is often characterized as a state of mind that is occurrent, one which the person is currently entertaining or which is currently occupying her mental attention. It has been alternatively characterized as a mental state that occurs in the person's stream of consciousness or which is held in her short term memory.[44] I will often refer to this distinction as the distinction between occurrent and non-occurrent mental states.

It is worth saying that there are potential differences in each of the above formulations: being occurrent, taking place in the person's stream of consciousness and holding in short term memory might not, ultimately, capture the same phenomena. Moreover, providing a precise and rigorous definition of any one of these distinctions is quite challenging. It is hard to distinguish clearly between certain "occurrent mental states" and certain "non occurrent" mental states. It is also difficult to draw a sharp line distinguishing between "short-term" and "long-term" memory as the former often fades into the latter. But even if a precise characterization might not be available and one can, at most, offer a vague characterization of the phenomenon, it is still possible to recognize, at least in broad strokes, what "occurrent mental state" is meant to encompass. This vague understanding will be sufficient for our purposes here.

In the dissertation I am arguing that to develop virtue it is important to come to *know* motives, desires, inclinations, beliefs, and other mental states that the ethical pilgrim does not know. Whether a mental state is occurrent or not does not allow one to distinguish whether one knows it or not. There are plenty of mental phenomena to which we are not paying attention, phenomena which are not occurrent, but which it seems natural to say that we know. Take the case in which I am explaining a difficult argument to my companion while driving a car. Because traffic is heavy and my lane is particularly slow, I switch to

44. Newell and David R. Shanks 2014; Baumeister and Bargh 2014.

another lane. But I do this "automatically," as psychologists would say. Thus, my switching to another lane is not occurrent. But even though my shifting to another lane might not have been in my stream of consciousness, it is still the case that I know why I was doing it; if my companion asks me, I can explain what I did and why, and this explanation would be a case of knowledge.

But even though the distinction between occurrent/non-occurrent states of mind does not correspond with the distinction between conscious/unconscious states of mind that I am drawing, there is an internal relationship between both distinctions. In particular, occurrent mental states stand to conscious states of mind in a necessary relationship: if a mental state is occurrent, then it is necessarily conscious. Conversely, if the mental state is unconscious, in my sense, then this state cannot be occurrent. Psychotherapists have developed certain terms to characterize the kinds of states of mind that are known by me but are not occurrent. They have used the terms "subconscious" or "preconscious" to refer to them.[45])

II.2.3 What Is Unconscious as What Is Repressed

Within psychotherapy, the unconscious is often characterized as the repository of the mental states that are repressed, kept outside of our conscious awareness because of motivational and emotional forces that make it hard for the person to acknowledge them. Psychologists usually highlight that their understanding of the unconscious is broader than Freud's. They argue that they are particularly interested in exploring those states of mind that are unconscious, not so much because of motivational factors but, rather, because of the architectonics of the mind.[46] This marks a significant difference between their investigation of the unconscious and the investigation of unconscious mental states that take place within psychotherapy. It is worth mentioning, however, that even if social psychologists think of unconscious states of

45. The latter term was introduced by Freud (Freud 1963, p. 296. The former term has been used, for instance, by Weintraub 1987, p. 424).

46. Timothy D. Wilson 2002, pp. 1–42.

mind mostly in computational terms, they deal with states of mind that involve motivational factors.[47]

II.2.4 Conscious/Unconscious Mental States Are Judgment-Sensitive

The kind of mental states that are at stake in the dispute between Annas, Arpaly and me are judgment-sensitive mental states, that is to say, mental states that could be directly formed, and are meant to be directly transformed, by the subject judging whether the mental state is one that it is merited to have.

It is natural to think that beliefs and intentions are judgment-sensitive. Arguably, some desires and emotions are as well. Judging or deliberating whether a desire or an emotion warranted is sometimes all that it takes for one to hold this desire or to develop such an emotion.[48] By contrast, passive states like sensations or pains are not supposed to be the result of, or to be transformed by, our judgments. Consequently, they are not judgment-sensitive.[49]

The expression "judgment-sensitive" can be heard as describing a generic type that sets up a norm or as particular species that instantiates a successful instance that complies with

47. Timothy D. Wilson 2002, p. 14.
It might be tempting to think that the set of repressed mental states is a subset of the set of states of mind that are unconscious in my sense. After all, a repressed mental state is one that the subject cannot self-ascribe. But this identification is not entirely correct. To see this, think of the case in which I come to know that I have a repressed mental state through the testimony of a psychologically acute friend. This knowledge will enable me to self-ascribe this repressed state of mind even while it is repressed. In cases like this, my mental state would be conscious (in my sense) but repressed.
In later chapters, I will refine my distinction between conscious and unconscious states of mind in such a way that, under this refined conception, the repressed will be a subset of the mental states that are not consciously held.

48. Saying that a certain emotion (or desire) is judgment-sensitive does not commit us to any form of extreme cognitivism nor to the claim that all types of emotions or desires (or even all instances of a particular type of emotion or desire) are judgment-sensitive. All that it means is that judgments ought to play a role in shaping and reshaping a number of our emotions and desires. If you ask: "which emotions and desires?" I respond "our *judgment-sensitive* emotions and desires."

49. I am borrowing the expression "judgment-sensitive" from the work of Richard Moran (Richard Moran 2001). He uses the expression to qualify desires that obey reasons, desires to contrast with brute desires like hunger or fatigue. In availing myself of this expression I am broadening its use beyond the case of desire.

that norm. In other words, one can call "judgment-sensitive" either a state that *could* be formed and *ought to be* transformed by the person's judgment or a mental state that is *actually* liable to be transformed by such judgment. In the dissertation I will use this expression in the first sense. Thus, to say that a mental state is judgment-sensitive does not imply that such mental state has actually been formed by a judgment nor that it is, in fact, sensitive to one. It only implies that it is the type of state that *could* be formed and *should* be transformed by such a judgment. Thus, a recalcitrant and irrational belief or intention that the person recognizes as such is "judgment-sensitive" because, even if it has not been formed, nor can be transformed, by the person's judgments, it is the type of state that, given its nature, should be responsive to these judgments. Saying that a judgment-sensitive mental state should be responsive to the person's judgments is like saying that a horse should have four legs. It does not exclude that there are three legged horses, but it does indicate that a three legged horse is a defective specimen of its kind. Sensations, in contrast to recalcitrant and irrational beliefs, are not the type of states that are meant to be formed or transformed by our own judgments. As a consequence, they are not judgment-sensitive.

The fact that a mental state is "judgment-sensitive" allows me to distinguish the set of unconscious mental states that I consider to be relevant for my project. It helps me to bring out that the unconscious mental states in which I will be interested are not subpersonal, that they are not associative aliefs and that they do not belong to what one might call the Lacanian unconscious.

Unconscious States of Mind Are Not Subpersonal

Many of the things that social psychology discusses under the rubric of "the unconscious" consist in sub-personal processes. Take the following remark from Wilson's book: "People cannot directly examine how many parts of their minds work, such as basic processes of perception, memory, and language comprehension, not because it would be anxiety provoking

to do so, but because these parts of the mind are inaccessible to conscious awareness—quite possibly because they evolved before consciousness did. If we were to ask people to tell us exactly how they perceive the world in three dimensions, for example, or how their minds are able to parse a continuous stream of noise emitted by another person into comprehensible speech, they would be quite tongue-tied."[50] These are processes that are executed by organs and systems in our body which do not belong to the realm of judgment-sensitive mental states. Because I am interested in those mental states that can formed and transformed by our judgments, I will not be interested in understanding subpersonal processes.

Unconscious States of Mind Are Not Aliefs

Tamar Gendler has coined the term "alief" to refer to a mental state that is "associative, automatic, and arational."[51] She appeals to this type of mental states to explain certain kinds of discordant behaviors and emotions, such as the reluctance of people to walk on the cantilevered glass walkway on the Grand Canyon, which they know is safe, or to drink from a bottle labeled "sodium cyanide" that they have seen has been filled with water and sugar. Gendler seems to suggest that aliefs are mental states that are not responsive to reasons but only to brute associations. If this is how aliefs are to be characterized, then it entails that aliefs are not *judgment-sensitive*. A judgment-sensitive mental state is a mental state that is supposed to be formed by and can be transformed by rational considerations. It is a mental state that is, at least potentially, responsive to judgment. Because aliefs are, by their nature, arational, they cannot be judgment-sensitive and, consequently, cannot be known first-personally.[52]

50. Timothy D. Wilson 2002, pp. 7-8.

51. Gendler 2008a, p. 641.

52. All that I am suggesting here is that the unconscious mental states in which I am interested are judgment-sensitive. If one wants to conceptualize (perhaps certain) aliefs as responsive to reasons, characterizing their lose associativity and recalcitrance to judgment as a certain kind of defect, then such aliefs

The Unconscious States of Mind I Am Talking About Do Not Belong to the "Lacanian Unconscious"

Within the psychoanalytic literature there are a number of authors, many of them influenced by Jacques Lacan or Melanie Klein, who believe that it is a mistake to think of repressed mental states which are at the heart of psychoanalysis as beliefs, intentions or emotions that happen to be outside of the person's awareness. Unconscious mental states, according to this view, have a structurally different nature from ordinary mental states. Consequently, they are not judgment-sensitive.[53] The kinds of mental happenings in which I am interested here, the desires, emotions or beliefs that are meant to rationally inform our behavior, are mental states that belong to the genus of our judgment-sensitive mental states.

II.2.5 The Unity Among Conscious and Unconscious States of Mind

Social psychologists sometimes describe the collections of conscious and unconscious mental states as constituting two entirely distinct systems, housed in different parts of the brain, operating according to different rules and having different roles in our cognitive economy.[54] This might be a good way to characterize, for instance, the difference between the activities that can be attributed to a person and those which are subpersonal. But it is inadequate to characterize the relationship between states of mind that are conscious and those that are unconscious. The kinds of unconscious states of mind which Arpaly discusses (and which, I am claiming, should be known by the person) have the potential of becoming conscious and, consequently, cannot be conceptualized as forming a collection of mental states that

would be part of my topic. Gendler, however, seems to suggest (and has been read as suggesting) that aliefs are not a defective specimen of the species belief, but rather that they constitute a different species.

53. For a philosophical account that pursues this line of thinking, see Gardner 1993.

54. See, for instance, Timothy D. Wilson 2002, Dijksterhuis and Nordgren 2006 or Baumeister and Bargh 2014.

are isolated in any principled way from the collection of mental states to which the person is consciously committed.[55]

This is not to deny that psychologists and psychotherapists are onto something when they attribute to unconscious mental states certain characteristic features that are different from characteristics features of conscious mental states. I will propose, however, that we should think of these interactions and manifestations as consequences of the vicissitudes that they undergo when the person is unconscious of them and not as essential properties of their being unconscious. Failing to do this, failing to keep in mind that there is no principled road block that prevents unconscious judgment-sensitive states of mind from becoming conscious, undermines the fact that all the judgment-sensitive mental states of a person are meant to form, at least in the ideal case, one single rational unity.

II.3 Proper Reasons Are Conscious Reasons

II.3.1 Conscious Mental States Tend to Be More Rational

With these distinctions under our belt we can now return to Arpaly's work. Arpaly claims that "[r]easons for action need not enter into consciousness to be excellent reasons or to justify actions (...) awareness of one's reasons need play no role in an evaluation others might correctly give of those reasons"[56] One of the aims of this section is to bring out the ways in which this assertion is problematic. To spoil the plot: I will argue that empirical

55. The proposal that conscious and unconscious states of mind form two different systems often leads those who favor such proposal to think of each of these set of states of mind as representing two distinct quasi-persons which are required, in light of the fact that they are housed within one single body, to negotiate and coordinate with each other so as to live harmoniously. This conception is problematic. It is certainly true that the coexistence of conscious and unconscious states of mind that conflict entails a division within the person. But this division needs to be thought as a division within *the* person, not a division between quasi-persons. When there are states of mind which conflict, one can see the person as divided. But she is divided, not into two sub-persons, but into competing or conflicting aspects of herself. For an illuminating discussion of this topic see Finkelstein 1999.

56. Arpaly 2003, p. 67.

research suggests that for a reason to be a proper reason, to be properly rational, it must be conscious. I will provide three distinct but interrelated arguments for this claim. I will argue 1) that unconscious processes tend to be more associative and less rational;[57] 2) that unconscious thinking is blind to negation and not sensitive to syntactic structures of language; 3) that the unconscious is particularly responsive to the here and now, lacking the ability to orient itself according to general principles that have a long term horizon.

These arguments are meant to show that if Huck never entertained his non-racist beliefs about Jim's humanity consciously, there is a *prima facie* reason to think that these beliefs will not be fully rational and, to the extent that they justify his behavior, that they are not, as Arpaly suggest, "excellent reasons."

Conscious Reasons, Logic and Rationality

It has been a common theme both within psychotherapy (particularly within psycho-dynamic psychotherapy) and within contemporary empirical psychology that when mental processes take place unconsciously, they tend to be more associative and less rational. This contrasts with the traditional treatment of the topic of self-deception. Within these treatments, mental states which the person denies having, mental states about which she is self-deceived, are often conceptualized as ordinary mental states that the person is merely incapable of acknowledging.[58] The philosophical tradition has also generally conceptualized unconscious mental states (and the literature has focused, quite heavily, on unconscious beliefs) as ordinary mental states that just happen to be unconscious. The clinical experience of psychoanalysts, however, suggests that this might not be the best way to think about these states

57. To say that these mental states tend to be more associative and less rational is not to say that they are aliefs. Aliefs are mental states that are associative and arational. Unconscious beliefs are, in virtue of being judgment-sensitive, meant to be rational and not associative even if their being unconscious tends to make them behave in less rational and more associative ways.

58. Joseph Butler 2006; Davidson 1982; Pears 1984.

of mind. In dealing with his patients, Freud came to recognize that when mental states are repressed they tend not to behave like ordinary mental states. In particular, he noticed that when a judgment-sensitive mental state is repressed the rational links of grounding and grounded that are meant to hold it together with other judgment-sensitive mental states weaken, becoming replaced by associative connections that are often irrational.[59]

There is a growing consensus among social psychologists and cognitive scientists that when states of mind are occurrent they are likely to hook up with other mental states more rationally.[60] A number of experiments have shown that performance on logical reasoning tests is heavily dependent on being able to perform these tests when one is fully attending to them, that is to say, when these tests are occurrent to consciousness.[61] Logical reasoning deteriorates sharply when one is preoccupied with other thoughts and improves when one can focus one's whole conscious attention on such reasoning.[62]

Dual-process theories of thinking have become extremely influential within psychology. So much so that, as DeWall et al. mention, they are now almost assumed as axiomatic.[63] These theories posit that human beings have two different systems to process information. According to the characterization of Wim De Neys, the first system is supposed to solve problems by relying on prior information while the second system is meant to use reasoning according to logical standards. In a well known experiment, De Neys asked participants to evaluate syllogistic arguments that, because of false premises, produced conclusions that conflicted with daily experience. Participants were asked to evaluate the validity of such syllogisms. Some had to do this under high cognitive load (they were asked to remember a complex visual pattern while they were assessing the validity of the argument), while others

59. Freud 1953b.

60. Baumeister and Bargh 2014; Baumeister, Masicampo, and Vohs 2011; Gawronski and Strack 2004.

61. Neys 2006; DeWall, Baumeister, and Masicampo 2008; Baumeister and Bargh 2014, pp. 42, 45, 46.

62. Baumeister, Vohs, and Masicampo 2014, p. 21.

63. DeWall, Baumeister, and Masicampo 2008, p. 629.

were allowed to judge the syllogisms under no cognitive load or under low cognitive load (they had to remember no pattern or the pattern that they had to remember was simple). The results showed that if the cognitive load increased, the capacity to successfully evaluate the validity of the arguments decreased.[64] In another experiment, participants where asked to complete a number of symbolic logic puzzles. Some participants where asked to solve the puzzles while listening to a song and counting the number of times that a recurrent word in the song came up. The other participants were allowed to solve the puzzles without attending to the song. Participants in the high cognitive load condition solved fewer logic problems correctly compared to participants without cognitive load.[65] The general conclusion that these experiments are meant to substantiate is that the "conscious, reflective processing system is vital for logical reasoning."[66][67]

Someone might object to the relevance of these experiments. The objector might argue that these experiments are concerned with unconscious processes and not with unconscious states of mind. The objector might actually add that, in fact, in these experiments, these unconscious processes lead to conscious mental states. In De Neys' experiment, for instance, all participants who had to assess the validity of syllogism, reached a conscious conclusion about such validity; their having a conscious (and occurrent) mental state did not depend on the process being conscious or unconscious. The objection brings out that the unconscious processes and unconscious states of mind cannot be just mapped to one another.

64. Neys 2006.

65. DeWall, Baumeister, and Masicampo 2008.

66. DeWall, Baumeister, and Masicampo 2008, p. 628.

67. It is worth mentioning one objection that could be raised against these experiments, which depend on accepting "the standard assumption that cognitive load mainly preempts conscious processing while allowing unconscious and automatic processes to proceed essentially unimpaired" (Baumeister, Masicampo, and Vohs 2011, p. 341). But, as DeWall, Baumeister, and Masicampo mention, one could challenge this by suggesting that cognitive load affects not only conscious processes but also unconscious processes. One could argue, for instance, that the decrease on the person's logical capacities under cognitive load could be explained by the effects that this increased cognitive load has on the capacity to process information unconsciously. There is plenty of theoretical and empirical evidence to suggest that this is not the case (see, for instance, DeWall, Baumeister, and Masicampo 2008).

There is, however, an internal connection between conscious/unconscious processes and conscious/unconscious states that allows me to substantiate that these experiments lend credence to the main claim of this section. Allow me to illustrate it with De Ney's experiment. It is certainly true that all participants reach a conscious (and occurrent) conclusion about the validity of the syllogism. Some of them reach this conclusion while being distracted while others reach it while paying their full attention to the evaluation of the syllogism. Arguably the process of assessing the syllogism (be it conscious or unconscious) involved intermediate steps, steps that involved the creation and transformation of a number of mental states. It is plausible to assume that when the participants were distracted, more of these states (and more of these steps) were unconscious. Conversely, to the extent that participants were able to pay their full attention to the evaluation of the syllogism, the intermediate steps they took towards their assessment and the mental states associated with them were likely to be conscious. And this shows that, to the extent that our mental states are conscious (in fact, occurrent) and hook up with other conscious (and occurrent) mental states, they will have a better chance of being connected to one another in proper logical relationships. Because unconscious mental states are mental states that I am not entertaining consciously—in fact, that I am *incapable* of entertaining them consciously—they will tend to link together with other mental states in lose associative relationships, they will be less likely to stand in relations of reason-giving with other mental states.

Negation, Syntax and Semantics

Other research projects within psychology have provided further evidence that when mental states are unconscious they tend not to stand in proper reason-giving relations with other mental states. Roland Deutsch, Bertram Gawronski and Fritz Strack designed a number of experiments which showed that the kind of fast automatic processes that tend to be identified

with unconscious processes are blind to negation.[68] According to Deutsch, Gawronski, and Strack (2006), the fact that these processes do not respond to negation suggests that they do not follow rule–based procedures.[69]

Freud's clinical experience led him to a similar claim. He suggested that the unconscious (in his sense) was blind to negation: "In analysis we never discover a 'no' in the unconscious"[70] If a mental process fails to reverse the truth value of a proposition when such proposition is negated, this suggests that the system is blind to, or at least not fully responsive to, the law of non-contradiction. Consequently, if a mental process fails to invert the evaluative truth value of a proposition when such proposition is negated, it is hard to see how this mental process can be characterized as rational. It follows that in order to properly engage in rule-based procedures, one needs to attend to these procedures; they need to be occurrent.

An additional set of empirical results provides further support for this from a different angle. Pascale Larigauderie, Daniel Gaonac'h and Natasha Lacroix developed some experiments that showed that cognitive load interfered with detecting syntactic and semantic errors but not with detecting typographical and spelling errors.[71] A couple of years later, Peter Gordon, Randall Hendrick and William H. Levine showed that cognitive load interfered more with reading comprehension of syntactically complex sentences than with simple ones.[72] Being able to parse syntax seems to be a capacity inseparable from the capacity to understand the logical relationships at play in complex ideas and sentences. The experiments suggest that in order to properly do this, to properly understand the logical relationships between complex ideas, one needs to be able to entertain these occurrently in conscious awareness.

68. Deutsch, Gawronski, and Strack 2006.

69. Deutsch, Gawronski, and Strack 2006, p. 402.

70. Freud 1925, p. 371. See also, Freud 1953b, p. 186.

71. Larigauderie, Gaonac'h, and Lacroix 1998.

72. Gordon, Hendrick, and W. H. Levine 2002.

Research has shown that being able to consciously entertain complex sentences and ideas is required to understand them and their syntax.[73] Ethical behavior involves complex ideas that require us to make subtle distinctions. It requires, for instance, tracking the difference between instrumental and final ends and distinguishing between intentional actions and accidental happenings. The empirical research suggests that if a person is living a life according to principles that she cannot verbalize, these principles will not be truly ethical. They will lack the complexity, subtlety and nuance of actions informed by true ethical considerations. Take the case of Huck. I suggested earlier that it is not at all clear that his attitudes towards Jim are those of someone who simply has non-racist beliefs. Because unconscious motives inform Huck's actions, it is not clear that these motives instantiates the general principle "a person of color is a human being; he is not a piece of property." As I suggested, empathy and compassion are likely the main forces that led Huck to help Jim escape. The fact that he was able to have empathy and compassion for a black man already shows inklings of non-racist attitudes. But although empathy and compassion tend to produce praiseworthy ethical actions, and they in fact produced them in this case, these emotions may also misfire if they are not guided by general ethical principles that warrant them. The fact that Huck could not verbalize his non-racist attitudes suggest that these were not fully fleshed out, that they did not amount to a proper ethical belief.

The Irrationality of Short-Termism

If we think that one aspect of rationality consists in being able to assess the world with a long term view that considers multiple perspectives and not with immediatist responses to the here and now, then the empirical disciplines we are considering provide us with a third way to argue that unconscious states of mind tend to be less rational than conscious

73. Baumeister, Masicampo, and Vohs 2011, p. 348.

ones. Most psychologists agree that one of the characteristics of unconscious processes are that they are effective in the here and now only. Bargh and Baumeister, for instance, have argued that "[u]nconscious processes seem to correspond to 'old brain' systems of dealing constantly with present-time dangers and needs."[74] They conclude that "the unconscious evolved primarily to produce adaptive responses in the immediate present."[75] This concern with the present can be related to what psychologists have identified as a tendency for unconscious processes to display impulsive behavior (guided by pleasure) over deliberated and controlled behavior (guided by well-being). Malte Friese, Wilhelm Hofmann and Michaela Wänke showed that, when under cognitive load, people tend to prefer foods that they find pleasant but not particularly healthy (chocolates) over foods that they consider healthy but not as pleasant (fruits). When participants were asked to choose edible articles while memorizing and rehearsing an eight-digit number, they tended to choose chocolates; when they were asked to do this while memorizing a single-digit number, they tended to prefer fruits.[76] Timothy Wilson has also suggested that the unconscious mental states are "not governed by accuracy and accessibility alone," adding that "people's judgments and interpretations are often guided by a quite different concern, namely the desire to view the world in the way that gives them the most pleasure—what can be called the "feel-good" criterion."[77] This view has striking similarities with Freud's insistence that unconscious mental states are responsive to the pleasure principle and not to the reality principle. According to Freud, the pleasure principle is the way in which the primary processes in the unconscious operate. He characterized these processes as short-term oriented, striving towards the goal of gaining immediate pleasure and keeping away any current unpleasantness.[78]

74. Baumeister and Bargh 2014, p. 43.

75. Baumeister and Bargh 2014, p. 47.

76. Friese, Hofmann, and Wänke 2008.

77. Timothy D. Wilson 2002, p. 38.

78. Freud 1958f.

Once again the argument here is an argument about unconscious processes and not about unconscious states of mind. But as I mentioned above, the conclusions about unconscious processes will also hold about unconscious states of mind: unconscious states of mind will tend to be responsive to the here and now, while conscious states of mind will have a better chance of being responsive to long-term considerations about how to live one's life as a whole.

How This Bears on Arpaly

Arpaly suggest that we can perform inferences unconsciously.[79] The clinical experience of psychotherapists coupled with the empirical research offered by psychologists suggests that although there might be certain kinds of simple reasoning tasks that we can do without consciously attending to them, and although there might be complex computations that are better performed through unconscious thought, full blown logical reasoning requires that we consciously attend to the process. Unconscious inferences tend not to behave in properly rational ways. Attending to these process, in turn, will tend to make the mental states involved in them hang together in more adequately reason-giving relationships with other mental states.

These experiments, of course, do not provide an unquestionable proof for my thesis, namely, that the rationality of one's states of mind improves when one is consciously aware of them. First, one could challenge their ecological validity. One might argue that asserting that mental states need to be conscious in order to stand in proper reason-giving relations warrants more than a few experiments evaluating the success of participants in a a narrowly

79. Arpaly 2003, p. 21. Although Arpaly clarifies that there is a sense in which deliberation requires the first person experience of distancing yourself from that on which you are deliberating, she nevertheless states that there are certain ways to understand "deliberation" according to which one can be said to be capable of deliberating unconsciously (Arpaly 2003, p. 22).
Versions of the idea that deliberation and inference can happen both consciously and unconsciously also been defended by a few empirical psychologists (See, for instance, Dijksterhuis, Bos, et al. 2006; Dijksterhuis and Nordgren 2006).

confined set of logical tests. In response one can point out that, although these experiments are certainly constrained to a narrow set of logical tests, they do assess central features that have to be in place in any reason-giving relationship: the capacity to identify that by negating a proposition one is reverting its truth value, the capacity to assess valid syllogisms and the capacity to track complex syntax. Coupling these experiments with the clinical experience of psychotherapists provides further support for my thesis. Thus, even if these experiments and clinical experiences do not provide unquestionable warrants to establish such thesis, they do provide a very strong case for it.

Second, the findings of social psychologists (and, one could argue, to an important extent also of psychotherapists) apply only to mental states that are mentally occurrent and not to the more general category of conscious mental states in which I am interested here. Therefore, someone might object that these results, which bear on occurrent mental states and not on conscious mental states, cannot be used to defend the claim that conscious states of mind, as a category, tend to be more rational than unconscious mental states. My response to this worry consists in biting the bullet (in fact, and as we will later see, it has important implications about how we understand the specific role that self-knowledge plays in ethical development).[80] But biting the bullet does not really undermine my thesis. The findings discussed so far support two facts. First, that the reason-giving relationships of unconscious mental states are less reliable. Because we cannot consciously attend to our unconscious mental states, the rational links of grounding and grounded that are meant to hold them together will tend to weaken. This is, already, an important insight that helps us to understand why we need to know our mental states: if they are unconscious, they will tend to behave in associative networks that will tend not to be rational. Second, even if the research bears only on occurrent mental states, we can connect the scope of their results

80. I will actually argue that ethical development requires, not just that we know our mental states, but that we exercise our ability to consciously know them, that we exercise our ability to occurrently entertain them.

with the more expansive category of unconscious mental states. One can do this by pointing out that there is an internal connection between being conscious of a mental state and the rationality of such a mental state. As the reader might remember, being conscious of a mental state, as I have defined it, consists in being able to self-ascribe such mental state. I mentioned that being conscious of a mental state is a capacity, the capacity to self-ascribe this mental state, a capacity that need not be exercised. When this capacity is exercised, however, the person who exercises it will be actually self-ascribing the mental state. And when she does this, the mental state will be occurrent. Thus, even if the previous results do not support the view that being conscious of a mental state tends to be a precondition of its rationality, it does support that the *exercise of this capacity* is a precondition of its rationality. Thus, even if these results do not prove that being conscious of your mental states contributes, in and of itself, to their standing in proper reason-giving relationships, it does prove that the exercise of this capacity will contribute to it. And this serves as a justification for why it will be valuable to develop such a capacity: only if one has it will one be able to exercise it.

II.3.2 Neil Levy's Account *vis-à-vis* Mine

A number of contemporary philosophers have taken notice of these results within cognitive science and have appealed to them when making philosophical interventions on some of the issues I have been discussing.[81] Neil Levy, in particular, has appealed to them to argue that agents need to be conscious of the reasons for which they act to be morally responsible for them.[82] Arpaly is, of course, one of his central targets. His response shares a number of features with mine; contrasting both accounts will help to sharpen my own.

It is standard for cognitive scientists to endorse dual process theories of cognition. These

81. See, for instance, Gendler 2008b, 2014, 2008a; J. Nagel 2014; Levy 2013, 2014; Hills 2015.

82. Levy 2013, 2014.

theories propose that human beings have two distinct cognitive systems to process information. The first is typically characterized by being fast, effortless, automatic, nonconscious, inflexible, heavily contextualized, and undemanding of working memory. This system is often identified with "the unconscious." The second, often times associated with what is consciously occurrent,[83] is meant to be slow, effortful, controlled, conscious, flexible, decontextualized, and quite demanding of working memory.[84]

Levy, like other philosophers influenced by this body of research, has followed these psychologists in conceiving our cognitive powers as sharply divided between these two systems. This has led him to insist that "conscious processes *alone* are rule-based."[85] "Deliberation," he claims, "requires the rule-based processing that is the domain of consciousness; the associative processes generated by unconscious attitudes should not be thought of as reasoning."[86]

At first glance, my view might look like Levy's. But there are some subtle but important differences between our views that are significant for the argument I am putting forth here. I will formulate these differences in the form of objections.

Let me start by saying that Levy is drawing out of the empirical literature more than what he is allowed to. As I said, the psychological experiments do not show that all unconscious mental states are irrational and associative. What they show is that our capacity to rationally engage with these mental states improves when those mental states are occurrent and tends to worsen when they are not. Levy, however, wants to argue that *only* conscious processes are rational. He defines conscious mental states as those that are "personally available." He clarifies: "Information is personally available when the agent is able to effortlessly retrieve

83. Timothy D. Wilson 2002; Levy 2013.

84. For some reviews of such theories, see Bargh 2007; Bargh and Morsella 2008; J. Evans 2008; Frankish and J. Evans 2009; Gawronski and Creighton 2013.

85. Levy 2013, 219. My emphasis.

86. Levy 2013, p. 222.

it for use in reasoning and it is occurrently online, actually guiding behaviour or mental processes."[87] It should be clear from this definition that Levy aims to encompass within his characterization of "conscious mental states," not only occurrent mental states but also states of mind that, though easily retrievable, can be at the moment outside of occurrent awareness. I argued before, however, that the empirical research does not show that personally available states are more rational, it only shows that a subset of these mental states, those that are occurrent, are more rational. Levy's claims about the rationality of consciousness, then, are not properly circumscribed to the specific set of mental states with which the experiments deal.

The reason to pursue this more conservative reading of the experiments is not merely one of parsimony and scholarly responsibility. I actually believe that we fail to properly characterize our unconscious mental life if we follow Levy's suggestion. In what follows I will put forth a couple of arguments that are meant to show why my account is preferable to Levy's. Although these are not knock-down arguments against levy's account, they should compel the reader to see why my account is more promising.

Levy draws a sharp division between the workings of conscious and unconscious mental states. But this sharp distinction does not seem to do justice to the fluidity between conscious and unconscious mental states. It is possible to make an unconscious state of mind conscious and occurrent. And it is also possible for occurrent mental states to become unconscious. The transition from transforming an unconscious mental state into a conscious one is, in fact, at the heart of psychodynamic psychotherapy. According to this therapeutic approach, when a conscious mental state generates too much anxiety the person represses it. Once it is repressed, the mental state can no longer be acknowledged by the person even though it keeps influencing and determining her behavior. Mental health is supposed to improve when repressed states of mind that are interfering with the person's life are made conscious. When

87. Levy 2013, p. 214.

they are made conscious, they tend to become more amenable and responsive to reason. And this, of course, presupposes that the nature of an unconscious mental state is such that it could become conscious, it presupposes that the nature of an unconscious mental state is such that, even if it does not stand to other mental states in proper reason-giving relationships, it is at least capable of doing so. If we think of an unconscious state of mind as a merely associative mental disposition, as Levy does,[88] it is not clear how they can become responsive to reasons in the way that they are meant to become when they become conscious.

Psychotherapy and literature provide us with many sources of examples that illustrate that our lives are pervaded by unconscious judgment-sensitive mental states (Arpaly draws heavily on both of them to bring this out). We often respond to unconscious motives which, even if we cannot verbalize, do *justify* our actions. We often experience unconscious emotions that we are unable to self-ascribe but which constitute *warranted* responses to our specific situation. I emphasized the words "justify" and "warranted" to flag the fact that these unconscious motives and emotions belong to the space of reasons. They form part of a network of states that are meant to be linked together by reason-giving relationships. Our unconscious motives, emotions or beliefs are judgment-sensitive even if they are not entirely rational, even if they do not stand with other judgment-sensitive mental states in proper reason-giving relations. Their rational flaws make them defective *qua judgment-sensitive* mental states, but these flaws do not disqualify them from belonging to the set of judgment-sensitive mental states. (To return to our previous analogy, a three legged horse is a horse even if horses are meant to have four legs). Considerations about the processing constraints on higher order cognitions or evolutionary facts about the development of such processes might explain why it is a good thing that we compute higher order information in associative ways. But the mere fact that a mental state tends to stand to other mental states in associative relationships that are not entirely rational does not, in and of itself, prove that

88. Levy 2013, p. 215.

this mental state is not judgment-sensitive.

We can formulate the difference between Levy and myself as a difference in how we think of the taxonomy of mental states. Levy and I are in agreement that conscious mental states belong to the genus "judgment-sensitive mental states." Levy, however, conceives of what we ordinary call unconscious mental states as belonging to a different genus, the genus of "dispositions which are partially constitutive of judgment-sensitive mental states."[89] It is as though, to return to Gendler's terminology, Levy is conceptualizing all unconscious beliefs as aliefs. My alternative suggestion is that we think of both conscious and unconscious beliefs as belonging to the same genus: "judgment-sensitive mental states." To the extent that unconscious beliefs tend to link together with other beliefs merely associatively, they will be defective specimens of its genus. As I have suggested, drawing too sharp a division between unconscious and conscious beliefs, conceiving of them as belonging to two different genera, deprives one of the resources to explain the fluidity between unconscious and conscious states of mind.

In reviewing Levy's work, Arpaly grants to Levy that on certain occasions some of our behaviors are the result of associative processes that are morally neutral.[90] But she highlights (and this is one of the points that I am trying to make) that there are certain kinds of prejudice that do not seem to be the mechanical result of associative process but which, rather, seem to be motivated by what we characterize as reasons.[91] My suggestion is that we follow Arpaly in thinking that in many of these cases such motivation is not a mechanical disposition produced by blind associations. It is normatively and rationally laden and, even if defective, should be considered to be judgment-sensitive.

All of these considerations are meant to provide the reader with some grounds about

89. My characterization of the genus that he would propose comes from what Levy writes about beliefs in Levy (2013, p. 215).

90. Arpaly 2015, p. 831.

91. Arpaly 2015, p. 831.

why we should prefer my conceptualization of unconscious mental states over Levy's. As I mentioned earlier, I would have to say more to establish this in a definite way. I would have to entertain Levy's responses to my challenges and the theoretical alternatives involved in these responses. It is beyond the limits of the project to do this.

My account, then, differs from Levy's in the way in which it characterizes unconscious mental states (and processes). Levy proposes that unconscious states of mind are associative *tout court* while conscious states of mind are not. I am proposing, first, that it is unwarranted to draw this sharp dichotomy from the empirical results and, second, that this sharp dichotomy fails to do justice to the nature and function of our unconscious mental life. What I am proposing is not, like Levy, that our unconscious states of mind are associative *tout court*, but that when we do not consciously attend to our mental states they *tend* to link together with other mental states more associatively and less rationally.[92]

One of the virtues of my account is that it allows for cases in which a person's unconscious mental states might be more rational than those which are the result of her conscious deliberation. Arpaly provides us with compelling examples to illustrate that this is a live option. Her examples help us to see that we often treat certain persons as good people who happen to be bad at abstract thinking or who allegedly hold unethical views to which they are not ultimately committed. "In these cases," writes Arpaly "we do not say, 'He's a bad person, but luckily he's weak-willed.' We say, 'He's a good person. Unfortunately, he has these silly views, but you can safely ignore them.'"[93] Withing Levy's proposal, it is hard to make sense of these cases.

92. There are moments where Levy comes close to formulating things in just this way. Take, for instance, the following remark: "unconscious information processing is *more associative* and *less rule-based* than conscious [information]" (Levy 2013, p. 219). His overall position, however, is one where conscious and unconscious mental processes are sharply distinguished as belonging to two genera.

93. Arpaly 2003, p. 9.

II.3.3 What Arpaly Gets Right and Wrong

My account is meant to do justice to Arpaly's examples in the way that levy's account cannot. At the same time, considerations that share many features with Levy's account are what enable me to put pressure on Arpaly's account. I side with Arpaly in thinking that there are persons who are, all things considered, more praiseworthy (or blameworthy) for what they do unconsciously than for what they do out of conscious deliberation. However, I do not think that these examples show, as she suggests, that the first person perspective is not important.[94]

Arpaly claims: "A theory of rationality should not assume that there is something special about an agent's best judgment. An agent's best judgment is just another belief, and for something to conflict with one's best judgment is nothing more dramatic than ordinary inconsistency in belief, or between beliefs and desires."[95] The empirical results that I have been discussing so far are meant to show us that that there *is*, in fact, something special about an agent's best judgment or, to be more specific, that there is something special about the agent's capacity to occurrently entertain her mental states.

Arpaly puts conscious and unconscious reasons on the same footing. But the clinical experience within psychotherapy coupled with the experiments in empirical psychology discussed above show that they should not be conceived in this way. Unconscious reasons tend to be less rational and conscious reasons tend to be more rational. As a consequence, there is a *prima facie* reason to identify rationality with our conscious activity and to think of unconscious activity as irrational (or, at least, as less rational). Reasons that one entertains occurrently tend to link with other judgment-sensitive mental states in reason-giving relationships that are more rational.[96]

94. Arpaly 2003, pp. 17-20, 22–3.

95. Arpaly 2003, p. 61.

96. I believe that this helps to 1) explain why philosophers tend to identify rationality with our best

I side with Arpaly in thinking that, in refusing to turn in Jim, Huck was acting from an unconscious moral motive. But I still hold that, because his moral motive is unconscious, it is not fully fledged. Although there is a lot of unconscious activity reshaping Huck's responses to racism, it seems plausible to think that, so long as Huck cannot articulate and reflect on his non-racist beliefs, these beliefs will not be fully formed. As a consequence, Huck's contribution to Jim's escape will not be grounded in a fully-fledged moral motive.

According to Arpaly, Huck is gradually coming to recognize Jim's humanity. And this recognition leads him to think that Jim should not be treated as a piece of property. My point is that it is problematic to think that this recognition should be counted as a fully-fledged non-racist belief. Arguably, part of what leads Huck not to turn in Jim has to do with the sympathy that he has developed for him. Arguably, this affective response will be part of what motivates Huck to refuse to turn in Jim. And to the extent that this is part of his motivation, it suggests that Huck is not acting in a non-racist way but that he is, rather, acting out of sympathy. One suspects that Huck's response would not be the same if the person involved was a slave that Huck had just met or one that he disliked. But for Huck's motive to be, properly, a non-racist motive, it would have to arise from the recognition of the humanity, not just of Jim, but of African Americans in general. And it is not at all clear that Huck holds this more encompassing belief. In fact, it is not really plausible to think that he does, given what we know of the way in which unconscious processes tend to work. The difficulty here is not merely epistemic, as Arpaly sometimes seems to suggest.[97] The issue is not merely that it is very difficult, or perhaps impossible, to know the specifics of Huck's unconscious beliefs. The problem here is that we have a *prima facie* reason to think that these unconscious beliefs will not be fully formed ethical reasons, that they will not have sufficient specificity and complexity to count as a properly fully-fledged non-racist belief.

judgment and, 2) suggest why this identification is warranted.

97. Arpaly 2003, pp. 60–1.

Moreover, the empirical literature suggests that unconscious beliefs tend to link together associatively and to be deeply intertwined with affective impulses that respond to short-term needs and not to long-term considerations. As such, these unconscious beliefs are not fully formed beliefs and should not be conceived as fully formed beliefs (although, as I was suggesting when I was criticizing Levy's position, they should still be conceived as beliefs).

II.4 Self-Knowledge and Ethical Development

My main aim in this chapter is to establish that knowing one's states of mind (or, to be more precise, knowing one's states of mind by being *consciously* in such states) is necessary to develop virtue; the person with ignorant virtue, I claim, is not fully virtuous. I've already indicated one reason for this. Empirical evidence suggests that the person with ignorant virtue is unlikely to act in ways that are fully rational. A person with ignorant virtue can be said to be acting for ethical reasons which she cannot articulate, but the fact that she cannot articulate these verbally suggests that these reasons will not be fully rational. In this section I would like to discuss a number of additional considerations that justify that developing and achieving virtue requires self-knowledge.

II.4.1 Conscious Deliberation and Rationality

Arpaly seems to acknowledge that there are ways in which it would be beneficial for the person with ignorant virtue to have knowing virtue. But although she accepts this fact, she seems not to recognize the full significance of it. Take the following claim, cited above: "A theory of rationality should not assume that there is something special about an agent's best judgment. An agent's best judgment is just another belief, and for something to conflict with one's best judgment is nothing more dramatic than ordinary inconsistency in belief,

or between beliefs and desires."[98] Formulated in the abstract, Arpaly's assessment seems correct. In the abstract, a reasons is a reason, regardless of whether it is conscious or unconscious. But when we consider the empirical facts of embodied human existence this abstract position becomes less plausible. Given what I have discussed from the empirical and clinical literature, we have a *prima facie* reason to think that an unconscious reason is less rational than a conscious reason. There *is* something special about an agent's best judgment. When an agent deploys her best judgment she does so occurrently, and this gives this form of deliberation a distinctive superiority to what one might call unconscious forms of deliberation.

Of course, it is better to be a person with ignorant virtue, acting well without knowing that one does, than to be a person with misguided conscience, believing that one is acting well while in fact being loyal to a cause that is morally irrelevant or downright evil.[99] On this I agree with Arpaly. Ignorant virtue is better than misguided conscience. But these two cases should not lead us to conclude, as Arpaly seems to, that conscious and unconscious reasons are on a par. As I said, all things being equal, the former is better than the latter. The fact that our conscious judgments might sometimes be misguided does not speak against their importance and the central place that they play in the development of a flourishing life. Take the case of Huck. If he could know his unconscious reasons and his unconscious motives he would be able to deploy them in his conscious deliberations and this would likely make the actions that follow from them more rational. Moreover because Huck's non-racist aspirations are unconscious, they will inform his actions without the intervention of his conscious deliberations. As such, the actions that would ensue are unlikely to be entirely rational.

98. Arpaly 2003, p. 61.

99. Arpaly 2003, p. 10.

II.4.2 Mediating Conflict.

When mental states conflict (say when you have desires which clash), being able to occurrently hold these conflicting states of mind is also said to be important for mediating and resolving the conflict.[100] "When the person has multiple motivations that produce competing, incompatible impulses, consciousness may help decide which one takes precedence."[101] This suggests that Huck might not be able to properly negotiate between his conflicting motivations because some of them are not available to his conscious attention. Huck faces a conflict between a desire to turn in Jim and a desire not to turn him in. His final determination not to turn him in, however, is not likely to emerge from a proper recognition of what speaks for and against each of these conflicting desires.. In fact, when we are faced with conflicting motivations, the proper response is often not to just arbitrarily pick one or another, but to find some sort of compromise where we are able to address what is warranted and to reject what is unwarranted in each of these conflicting motivations. This kind of nuanced response to conflicting motivations seems to be unavailable for a person like Huck, faced with motivations of which he is ignorant.

II.4.3 Disunity

The person of ignorant virtue is someone who is unconsciously guided by a supposedly virtuous orientation towards life but who is not consciously aware of such an orientation. But this will entail that she will be disunified in important ways. Because her conscious and unconscious goals will not be the same, she will be constantly sabotaging herself. Whenever unconscious goals conflict with her conscious goals, the pursuit of one set of goals will conflict with the pursuit of the other and *vice-versa*. Arpaly concedes a version of this point. "Huckleberry Finn" she writes "is not a bad boy who has accidentally done something good,

100. Baumeister and Bargh 2014, p. 44.

101. Baumeister, Masicampo, and Vohs 2011, p. 351.

but a good boy. No doubt he is imperfect, and one who would be better if some of his moral convictions were changed, but as he is, he is better than many."[102] What I want to highlight, that she does not, is that his conflicting motivations and reasons are going to have implications for how he acts. And they are unlikely to always run in the right direction. In the "golden opportunity" to turn in Jim, which Arpaly discusses, Huck's unconscious reasons trumped his conscious ones, leading Huck to do the right thing. But it is unlikely that this will always happen. As I have suggested, if Huck was faced with the opportunity to turn in an escaped slave for whom he felt little sympathy, it is likely that his conscious motivations would trump his unconscious ones. Furthermore it seems plausible to think that Huck will not act in a reliable way with respect to his non-racist disposition. A person's commitment to a certain view disposes him to act in a coherent set of ways. We expect a person who values the equality races, not merely to comport with respect towards people of color, but to defend this view when it is attacked and be upset when it is flouted. There is little evidence to think that Huck displayed this behavior. As such, Huck is someone who is constantly sabotaging this aspect of himself that rejects racism, undoing with some deeds what he attempts to achieve with others.

There is a further problem with the disunity at play in the person with ignorant virtue. A growing body of literature shows that people whose conscious and unconscious goals are more congruent tend to have greater emotional well-being. "[L]acking congruence between implicit and explicit levels of motivation is associated with impaired life satisfaction and emotional well-being, increased psychosomatic complaints and medication use, and clinically relevant levels of depressive symptoms"[103] If this research is correct, it would suggest that the disunity I was just mentioning would tend to lead Huck, not merely to self-sabotage, but also to a deterioration of his mental health. If one identifies the criteria by which

102. Arpaly 2003, pp. 77–8.

103. Schultheiss and Strasser 2012, p. 40. See also: Timothy D. Wilson and Dunn 2004; Baumann, Kaschel, and Kuhl 2005; Hofer, Chasiotis, and Campos 2006; Schultheiss and Strasser 2012.

psychologists measure mental health with flourishing, this entails that the disunity in the person with ignorant virtue will impair her capacity to live a properly flourishing life.

II.4.4 Planning, Simulating and Reflecting

Empirical research suggests that there are a number of capacities that are not available to the person with ignorant virtue. Being conscious of one's states of mind is necessary to deploy these states of mind to make long term plans and simulate the future. Similarly, conscious reflection on the effectiveness of one's actions has been shown to improve one's performance.

Most contemporary psychologists working in the area agree that making long-term plans requires that we are able to consciously entertain them.[104] Bargh and Baumeister claim that "[f]or all its powers and merits, the unconscious is probably not capable of making a complex plan itself, so it uses consciousness in order to make the plan."[105] Interestingly enough, the clinical experience of psychotherapists gives further support to this claim. Within the consulting room, repressed states of mind mostly manifest through outbursts and disruptions, not through well-crafted intentionally actions.[106]

A central part of making a plan consists in one's capacity to imagine how the plan plays out into the future. We can imagine and simulate different courses of action and their probable outcomes which can help us make decisions about how to act. Research in psychology, however, suggests that these kinds of simulations into the future require that one is able to consciously entertain these potential actions or states of affairs.[107] These simulations facilitate decision making, allowing us to choose and decide how to react to a particular situation. This is significant from an ethical point of view. As Darcia Narvaez

104. Baumeister and Bargh 2014, p. 43. See also: Baumeister, Masicampo, and Vohs 2011, p. 336.

105. Baumeister and Bargh 2014, p. 43.

106. If these unconscious mental states manifest in something like long-term goals, premeditated planning, or fully-fledged intentional actions, it is by influencing conscious mental states.

107. Baumeister and Bargh 2014, p. 44.

and Kellen Mrkva have argued (and it is a remark which I will revisit shortly): "The moral life involves co-authoring the future with others through dialogue and feedback on imagined alternatives."[108]

It has also been shown that conscious reflection on feedback or outcomes can improve the outcomes that one is seeking to bring about.[109]

Someone might object that what I have just been saying does not really apply to the person with ignorant virtue. Think of the case of Huck. Huck, who is meant to possesses ignorant virtue, does the right thing without being able to articulate why he does so. But Huck is entirely aware of his intentions; he is consciously aware that he wants to help Jim escape. And this, the objector would argue, is all that he needs to be able to make long-term plans about his decision, simulate the future so as to bring about those plans, and to reflect on his performance (and on the feedback he got, if any).

The objection fails to recognize, however, that even if Huck is aware of his *specific* intention to help Jim, it is not entirely right to say that Huck is aware of his *overall* intention. It is certainly true that Huck consciously knows he wants to help Jim escape but, because his non-racism is unconscious, there are aspects of his intention that will not be available to him. For instance, Huck will not be able to make long term plans to transform the racist conceptions that the people have about Jim, to prevent certain types of racist aggressions against Jim or to instigate and help other slaves escape.

Having one's virtuous orientation consciously available is not merely valuable to allow one to deploy it in (conscious) deliberation. In a review article mentioned above, Baumesiter, Masicampo and Vohs mention that a robust body of evidence has shown that "imagining oneself doing something can increase the likelihood or efficacy of doing it, especially in the

108. Narvaez and Mrkva 2014, p. 26.

109. Baumeister, Masicampo, and Vohs 2011, p. 339. It is worth mentioning that such improvements occurred when the person *both* reflected on these issues *and* was given feedback. Reflection without feedback brought no benefit (. . .) Feedback without reflection was likewise unhelpful (Baumeister, Masicampo, and Vohs 2011, p. 339).

future."[110] This suggests that being able to have one's virtuous orientation present to mind, in fact, being able to imagine oneself living the life that one aspires to live, is likely to help one actually succeed living such a life. Huck, however, is deprived of this possibility. Being unable to self-ascribe his non-racist beliefs makes it unlikely that he will imagine himself comporting himself in non-racist ways. In fact, because he consciously avows his racist beliefs, it is more likely that he will imagine himself, in the future, replicating them. And this, according to the empirical research, will make him more likely to act in racist ways than in non-racist ways.

II.5 Speaking One's Mind and Ethical Development

Nearly all psychologists working in this area agree that occurrent thought is required for verbal communication. Even though there is evidence that single words can be processed unconsciously (that is to say, that single words can elicit behaviors that indicate responses to the spoken words), properly parsing and understanding fully formed sentences requires occurrent mental thinking.[111] This, of course, does not mean that all communication needs to be mentally occurrent, nor is it meant to deny that there are all kinds of unconscious aspects (even unconscious communicative aspects) to verbal communication. It is just to say that it is essential to my capacity to communicate verbally that I am consciously attending to what I am saying.[112]

110. Baumeister, Masicampo, and Vohs 2011, p. 335.

111. See, for instance, Baumeister, Vohs, and Masicampo 2014, p. 21, Baumeister and Bargh 2014, p. 40 Gendler 2008a, p. 649 Timothy D. Wilson 2002, p. 65.

112. Wilson argues that "people are even able to understand and retain some of what occurs when they are under general anesthesia. When patients are given suggestions during surgery that they will recover quickly, they subsequently spend less time in the hospital than patients not given the suggestions, despite having no conscious memory of what was said while they were under anesthesia" (Timothy D. Wilson 2002, p. 25.).

Why is this significant for ethical development? There are basically three sets of reasons that explain why articulating these mental states verbally contributes to such development. First, because it makes it more likely that the person's mental states (which include the reasons that ground them) will be more rational. Second, because being able to articulate these mental states (and reasons) is necessary for this person to be properly self-directed, something that any account of virtue should hope to promote. Third, because it allows this person to discuss these mental states with others and, in doing so, to transform these mental states and the reasons that ground them in response to these discussions. Allow me to clarify each of these.

II.5.1 Speaking (and Writing) About Thoughts and Emotions

When I have talked about the person of ignorant virtue in this chapter, I have mainly focused on a person who acts rightly for the right motives but who is unaware of his motives. I have focused mostly in the way in which knowing her reasons and motives can help her become virtuous. But it is worth pointing out that there are other kinds of mental states which she should be conscious of. James Pennebaker's research program has shown that a person's mental health improves when she speaks (or writes) about thoughts and emotions related with traumatic experiences that she has experienced.[113] "When people transform their feelings and thoughts about personally upsetting experiences into language, their physical and mental health often improves."[114] These improvements have been measured by fewer visits to physicians, fewer self-reported illnesses and less self-reported aspirin consumption over subsequent months. Academic test performance has also improved when students reflect on

This suggests that, at least in some level, unconscious states of mind are able to respond to speech. However, what research suggests is that these states of mind can be formed and transformed by speech, not because of the full-fledged meanings communicated through speech, but because of associative connections elicited by some of the spoken words.

113. Pennebaker and Chung 2014, p. 418.

114. Pennebaker and Chung 2014, p. 418.

their emotions.[115] If we think that these improvements in mental health and well-being, as measured through the standards used in these studies, are constitutive of a flourishing life, then it follows that examining her mental life and articulating it verbally will be important for the ethical pilgrim.[116]

It is through our capacity to speak that we articulate our ethical responses. This articulation is often, in an of itself, an ethical achievement. In formulating these responses in language we are often transforming many of our responses from inchoate intuitions that guide us, either unconsciously or unreflectively, into properly rational responses, embedded in properly reason-giving relationships. Empirical research has shown that verbally labeling an emotion influences the emotional experience.[117] Pennebaker has, in fact, hypothesized that "if an emotion or experience remains in analog form [that is to say, if it is not articulated verbally], it cannot be understood or conceptually tied to the meaning of an event. (. . .) Once an experience is translated into language, however, it can be processed in a conceptual manner. In language format, the individual can assign meaning, coherence, and structure. This would allow for the event to be assimilated and, ultimately, resolved and/or forgotten, thereby alleviating the maladaptive effects of incomplete emotional processing on health."[118] Not surprisingly, therapists have remarks that also support this view. In her influential textbook on Melanie Klein's psychoanalytic approach, Hanna Segal describes a case study of a patient referred to as Ann: "the real help I was able to give her was in naming the different feelings inside her helping her to know them, to differentiate them and, therefore, to feel more able to control them."[119]

115. Baumeister, Masicampo, and Vohs 2011, p. 338.

116. It might be relevant to mention that the research suggests that *how* one thinks about these traumatic events is important. Merely rehearsing and reliving the event or merely ruminating on it usually prolongs the unpleasant aspects of the experience rather than diminish. One needs, instead, to explore it, analyze it and investigate it. (Baumeister and Bargh 2014, p. 338)

117. Pennebaker and Chung 2014, p. 16.

118. Pennebaker and Chung 2014, p. 17.

119. Segal 1974.

Speaking About an Emotion Is Necessary to Evaluate it Normatively

Moreover, our capacity to speak about our beliefs allows us to engage with these beliefs in ways that are not available to someone who cannot speak about them. In their seminal paper, "The Self-Regulating Mind,"[120] Victoria McGeer and Philip Pettit discuss the difference between a subject who can express the contents of their beliefs and a subject who cannot. They argue that although both subjects might have beliefs, the first is blind to the contents believed but the second will be capable of paying attention to these contents. This capacity underpins a capacity to evaluate the nature of these contents and to make sure that these beliefs conform to these evaluations. As such, it provides the person with the capacity to be self-directed. An influential program within empirical research has reached similar conclusions: "Once an experience is translated into language it can be processed in a conceptual manner. In language format, the individual can assign meaning, coherence, and structure."[121]

This capacity to evaluate the nature of her mental states, to assign meaning, coherence and structure to it, will only be available to the person with knowing virtue. Verbalizing her mental states and the reasons that ground them allows her to embed these within the broader set of ethical principles to which she is committed. If she finds that they do not cohere well with these principles, this will lead her to revise either these mental states or her ethical principles. This is something that is not really available to the person with ignorant virtue. It is true that the mental states of the person with ignorant virtue might transform as a result of something akin to a failure of conformity with her ethical principles. Huck is a good example of this. He had been raised with deeply racist views, but his interactions with Jim gradually lead him to interact differently with Jim in response to what seems a

120. McGeer and Pettit 2002.

121. Pennebaker and Chung 2014, p. 17.

recognition (even if inchoate) of Jim's humanity. This recognition involved a transformation of all kinds of beliefs about black people (or at least about this black man). But what is worthwhile to emphasize is that these modifications were, to a significant extent, associative; they were responding to features of the environment that were statistically significant, but they were not responding to reasons as such. And even if these associative transformations can be said to lay the foundation for what can eventually become a non-racist belief, they are not yet a full blown non-racist belief.

Discussing and Teaching About Our Ethical Outlook Is Central to Ethics

Finally, articulating these responses in language is valuable because it allows the person with knowing virtue to discuss them with others and to evaluate them in light of how others respond. Talking about our motivations, desires, beliefs and emotions, and about the reasons that ground them, is a very significant source of ethical growth. Interlocutors who have different perspectives from our own will often help us identify blind spots and biased perspectives and, in doing so, will help us to shape our minds in such a way that we are properly responsive to ethical considerations.

We are social creatures who grow through our interactions. And although our non-verbal interactions are very important in our social life, what makes our life a truly human life is the fact that we interact with one another verbally. Verbal interactions play a unique role in helping us reshape and reorient our lives and those of others. As Narvaez and Mrkva have argued: "The moral life involves co-authoring the future with others through dialogue and feedback on imagined alternatives."[122]

The person with ignorant virtue is someone who cannot articulate verbally his ethical responses to the world. If he cannot articulate them, it will not be possible for him to have

122. Narvaez and Mrkva 2014, p. 26.

discussions about it and to grow and modify his ethical stance through these discussions. Furthermore, this person will not be able to teach someone to be virtuous in the way that he is. This is, of course, not to say that the only way in which we are transformed ethically is through these interactions. It is to say that there is something distinctive about this way to be transformed, something which is at the heart of what it is to be a human being.

As Allison Hills points out, elaborating an undeveloped remark made by Annas in *Intelligent virtue*, the importance of this is not merely that lacking the capacity to self-ascribe these mental states and the reasons that ground them deprives the person with ignorant virtue of these opportunities for ethical growth.[123] Hills suggests that most accounts of virtue consider that these activities are, in and of themselves, an essential part of human excellence. Helping others grow ethically by correcting their views or by educating them into what an excellent life looks like seems to be part being an excellent human being.[124] It is also characteristic of human life that we can provide justifications for our actions, justifications that articulate their moral significance. Hills makes the interesting suggestion that this is actually part of the virtue of justice: "The exchange of justifications (offering and accepting or rejecting them) is a way of relating to one another of fundamental moral importance. It is essential to people's interests as rational agents with moral standing. Recognizing and responding to this is part of having the right moral attitudes toward other people; giving people what you owe to them, giving them what they deserve, is part of the virtue of justice. Justice requires that [the virtuous person is] able to explain her decision."[125]

Allow me to return to Huck. Huck cannot self-ascribe the thought "Black people are human beings that should not be treated like property." Because he cannot verbalize this thought, he cannot argue for it and change other people's mind about it. Knowing it, on the contrary, would allow Huck to be more intentional about helping Jim, not only by trying to

123. Hills 2015, p. 27.

124. Hills 2015.

125. Hills 2015, p. 29.

change other people's attitudes about Jim but also by intentionally seeking strategies and policies that would help to promote this view (he could, perhaps, promote legislation that banned slavery and vote for it).

II.6 Summing Up

Let me sum up what I have been saying in the previous three sections. My main claim is that there are a number of tools and opportunities for ethical development that are available to the person with knowing virtue but not to the person with ignorant virtue. The virtuous motives of the latter will not figure in her conscious deliberations, making these deliberations less rational. Some of these virtuous motives will also be unavailable in crafting long-term plans, simulating a future where she is acting on these motives, and improving the adequacy in which these motives play out in her life. She will be unable to consciously imagine herself fulfilling her virtuous aspirations, and to use these aspirations to organize and unify mental conflicts that might arise or which manifest only unconsciously. Ethical development is guided by general principles that often give primacy to long-term goals over the present satisfaction of lower desires or impulses, and we have reason to believe that the person with ignorant virtue will often end up privileging the latter over the former. Moreover, being disunified about what she believes consciously and unconsciously will inevitably lead her to sabotage herself; her actions will frequently obey to motives that she consciously acknowledges but which undermine her unconscious motives (and *vice-versa*). These conflicts will take a toll on her mental health and well-being, something that will be worsened by her incapacity to talk or write about them. The person with ignorant virtue will not be as capable of integrating her conflicting mental states within a perspective that can envisage her life as a whole, not merely in the here and now. Finally, because the person of ignorant virtue will be unable to formulate her own mental states in words, her unconscious virtuous

mental states are likely to lack the proper logical complexity that characterize virtuous mental states. The person with ignorant virtue will also be unable to properly measure her actions in light of her ideals and, consequently, she will not be a properly self-directed agent. She will also be unable to have discussions about these mental states with others, depriving herself of the ability to be transformed by these discussions and to transform others through them.

II.6.1 The Power of Unconscious Thinking

I mentioned earlier that there are a number of perspectives from which one can say that our unconscious activities are smarter than our conscious activities. What I have been suggesting need not undermine this. It is possible to grant that some of our mental activities take place faster, more efficiently, and perhaps even more reliably, when they take place unconsciously; that unconscious processes can sometimes lead to better outcomes when we don't think about them occurrently; and that human beings can perform highly efficient and powerful computational processes unconsciously. It is possible to grant all of this and still hold that conscious thought is superior to the extent that it helps to promote more rational connections among mental states. This entails recognizing that the kinds of activities that one might label "unconscious thinking" are, for all their computational power, not best described as rational thinking. In some instances, the power of unconscious thought might be deployed in the service of making what might resemble rational thinking. But the tendency for unconscious thought to operate associatively should help us to see that, at least in the paradigmatic cases, unconscious thinking is not rational. There are plenty of activities of thought that have been called unconscious decisions or unconscious deliberations. If what I have written in this section is correct, however, we should see that, at least in the standard case, these decisions or deliberations cannot be conceived as belonging to rational thinking but rather as calculations that receive certain inputs and produce certain outputs, calculations that

might resemble rational decisions or deliberations, but which, at the end of the day, should not be considered to be so.

The person with ignorant virtue relies on cues from the environment that are statistically significant. But she is unlikely to respond to proper reasons. And this will lead her either to fail to identify exceptions to an alleged pattern or to make proper generalizations based on it. In this respect, Annas was on the right track when she claimed: "because the person [with ignorant virtue] lacks the ability to demand, and give, reasons for what he does, he is not equipped to deal with new and unforeseen circumstances."[126] I say "on the right track" and not "correct" because Annas fails to recognize that the person with ignorant virtue is, in fact, equipped to deal with new and unforeseen circumstances. As I mentioned, human beings have the capacity to respond in very intelligent and flexible ways even if they might not be able to articulate them. Despite this, Annas is "on the right track" because this person's actions are guided more by associative connections than by properly reason-giving considerations. As Annas suggests, there will be occasions where these associations will lead this person to establish superficial similarities that will lead her to act wrongly: "[T]he 'naturally brave' person may get into a serious fight over a slight intended as a joke, while the 'naturally sympathetic' person may find herself the victim of scams."[127] Annas mentions that, according to Aristotle, those possessing natural virtue are like blind people who stumble and fall over. One needs to qualify this if one is speaking about the person with ignorant virtue. He does posses vision albeit an incomplete and distorted one. He can respond to features of the environment and track similarities and patterns that often lead to the right action. But lacking an understanding of the ethical principles that undergird these actions, he will sometimes fail to recognize that some of these similarities and patterns might be superficial and do not warrant such action.

126. Annas 2011, p. 25.

127. Annas 2011, p. 25.

In this regard it is relevant to mention that, despite all the alleged power of our unconscious to compute and find correlations, a good scientist will never feel satisfied with their gut feelings about a particular theory or empirical claim. Regardless of the power of the scientist's unconscious computational processes, these processes are often fallible. And it is only through careful conscious rational reflection that she can properly assess the validity and robustness of the insights that these process might bring up. I mentioned that it as an important part of ethical behavior to be able to guide and teach others. Something similar holds for scientists. A very important part of being a good scientist consists in convincing other scientists argumentatively that their hypotheses are true. And to do this they require to be conscious of what they are putting forth.[128]

II.6.2 Returning to Arpaly

I agree with Arpaly that some people who do not know themselves well are better (and might live more flourishing lives) than other people who do know themselves well; Ignorant virtue is better than knowing viciousness and better than misguided conscience. I am also willing to grant that we can attribute some level of virtue to people with ignorant virtue. Provided that we recognize that the person of ignorant virtue will not be properly virtuous; to be properly virtuous one needs to possess knowing virtue.[129]

I am also willing to grant that there might be individuals for whom it is better not to know themselves, individuals whose rational occurrent processes, for one reason or another, consistently malfunction. People whose power of reason is, say, systematically deployed to rationalize vicious inclinations. Granting this, however, should not preclude us from

128. It is worth mentioning, however, that if the research in empirical psychology is right, and our unconscious thought is capable of identifying certain patterns and correlations quicker than our conscious thought, then we might need to rethink how we are training our scientists, attempting to help sharpen these unconscious abilities so as to exploit them more intentionally.

129. My position is, therefore, similar to the position defended by Dougherty 2000, p. 123 and Hills 2015, p. 33.

recognizing that these kinds of individuals are exceptions and that, all things considered, it is better for a person to come to know her own mind and, more importantly, that self-knowledge is required to be properly virtuous.

One of the central ideas that I defended in this chapter is that the capacity to (genuinely) self-ascribe a mental state is necessary for the ethical pilgrim. Our capacity to speak is at the center of our capacity to reflect, evaluate and deliberate; it is at the center of our capacity to reason. It is through speaking that we can articulate what we take to be the response of the question "how should I live?" and that we can confront this response with alternative responses, challenge them, and get challenged about them.

When a person is able to come to know her unconscious mental states and exercise her capacity to hold them occurrently, these unconscious states will undergo a transformation towards increased rationality. As such, these mental states will become more amenable to interact according to the rules of rational thinking.

Of course, none of this is to say that conscious thinking is impeccable, that it always responds to the right reasons, or that it always performs fully rational operations. There are many ways in which our capacity to reason can be interfered with. But I have argued that our capacity to reason will be interfered with when we lack knowledge of our mental states. In other words, there are all kinds of things that can interfere with our capacity to live our lives rationally, holding unconscious judgment-sensitive mental states is one of them. Making them conscious can help us to live more rationally.

Psychologists are likely to reply that I am painting an unrealistic picture. It is not possible to avoid processing information unconsciously. Moreover, the amount of information that we can process consciously is but a fraction of the total capacity of the cognitive system to process information. Psychotherapists will also argue that the unconscious influences are massive and it is a fantasy to think that we can even make all of them conscious.[130]

130. Dijksterhuis and Nordgren 2006, p. 97.

My suggestion here is not that we should obsessively make everything that is unconscious conscious. I don't think that we can. My point is just that we *can* develop self-knowledge of some of our mental states and, thereby, to make these mental states more amenable to rational reflection. If these mental states are central in our lives, making these conscious is will likely have important ethical implications.

It is unclear to me whether Arpaly would agree on the position I have put forth here. Although I have defended statements that explicitly conflict with claims defended by Arpaly, some concluding remarks in her work seem to suggest that she agrees with my view.[131] Take this one: "Intellectual knowledge of morality and the ability to deliberate well about all matters moral have obvious things to recommend them. To begin with, an agent in Huckleberry Finn's position, who does not know his virtues from his vices, is likely to try to make himself a worse person, whereas the person who knows his virtues from his vices is likely to try to make himself a better person. More important, there are certain types of morally desirable actions that are very hard to perform if one does not have the right moral principles or at least the ability to deliberate well on moral matters. For example, it is very hard to vote for the right political candidate. (. . .) The fact that a person —even a good person—cannot deliberate very well or is ignorant when it comes to morally relevant matters has a special sadness to it."[132]

Arpaly mentions that these concluding remarks may make one wonder if she is not taking back some what she said before.[133] I wonder this myself. But clarifying whether these statements are inconsistent with the main things that she defends is not important for my purposes. My main aim is not to bring out inconsistencies in Arpaly's project but to show

131. I refer to the epilogue of chapter 2 and to the last few paragraphs in chapter 4 in *Unprincipled Virtue* as well as to the concluding section in "On Acting Rationally against One's Best Judgment," an article on which chapter 2 was modeled (Arpaly 2000).

132. Arpaly 2003, pp. 114–5.

133. Arpaly 2003, p. 65.

what is problematic about thinking that a person with ignorant virtue can be fully virtuous. Arpaly's work provides the best account I know of from which to defend this position. And it is against such position that I have argued against here. If Arpaly ultimately agrees with it, so much the better.

II.7 A Potential Objection to My Thesis

I'd like to close this chapter by discussing a potential objection that one can raise against the thesis that I have defended on this chapter.

Research on associative learning, priming influences, and automatic processes has shown that our behavior is shaped by myriad unconscious influences.[134] This research has conclusively established that the number of unconscious influences that guide or shape our behavior is almost boundless. Because these influences are so vast some scholars have suggested that the project of knowing oneself is hopeless. There is too much to know and it is, simply, not possible to come to know all of it.

This objection fails to properly understand the scope of the ethical project of knowing oneself. This project, rightly understood, does not consist in coming to know every minute influence that guides each corner of our behavior. It consists in coming to know those influences that are central to our lives. Keeping in mind the temporality of a life, often discussed under the guise of "character" by virtue ethicists, helps one to see that if these influences are central to our lives, then they will manifest in many of the activities in which we engage. In other words, the injunction "Know thyself!" is an injunction to come to know those things which keep coming up, those things which are *systematically* interfering with our capacity to live the life we find worth living. But it is precisely because these things keep coming up that they manifest pervasively in our lives. And this makes it feasible to aspire to

134. Timothy D. Wilson 2002; Hassin, Uleman, and Bargh 2005; Bargh and Morsella 2008J. M. Doris 2009.

know them. This is, of course, not to say that it is easy to acknowledge these repetitions or schemata; there are often huge emotional and cognitive barriers to our acknowledging them. My point here is just that the difficulty of coming to know these influences does not have to do with their abundance, with the fact that there are too many of them, but with other factors.

What I am saying is at the heart to a number of different schools of psychotherapy. Freud argued that repetition was one of the most crucial concepts in psychoanalysis. He argued that the client was meant to remember, i.e. to articulate verbally, the patterns of thought and action which she systematically repeated. It was these repetitions, which were acted out, that the client was meant to worked through.[135] Within Cognitive Behavioral Therapy (CBT), a form of mental treatment that has been often taken to be antipodal to classic psychoanalytic practice, therapists have similar aspirations. Identifying and changing schemata is at the heart of the treatment within CBT. Schemata represent organized patterns of thought and behavior that are used by the person to process information. These patterns of thought are said to structure the person's understanding of the world and, as a consequence, they keep coming up.[136,137]

135. Freud 1980.

136. A. T. Beck 1976; Reinecke and Freeman 2003; Kellog and Young 2008; J. S. Beck 2011

137. A somewhat similar objection to this arises from the fact that it is not possible to know whether one has gotten to the bottom of one's motives. Some authors, like Onora O'neil, have argued that the fact that we can never be sure of what our motivations are entails that the project of acquiring self-knowledge is hopeless (O'neill 1998). Janine Grenberg provides what I take to be a sound response to this challenge. She argues that the fact that our actions are often clouded by self-deception and misplaced self-love should not lead us to deny that self-knowledge is important or necessary for the development of virtue. "The agent needs to take with her into action the continuing question of who she is, why she has done what she's done —in Kantian language, what maxims she claims as her own— for this fallible knowledge will prove crucial to her in her ongoing pursuit of virtuous action. Over time, her confidence in who she is and why she acts as she does can grow, or be challenged, through reflection on her patterns of action" (Grenberg 2005, p. 220).

III

First/Third-Personal Self-Examination

In the previous chapter I established that ethical development requires self-knowledge. The ethical pilgrim, the person seeking to become virtuous, should come to know her mental states and the reasons that ground them.

The main question that I intend to address in the rest of the dissertation is whether it matters how the ethical pilgrim comes to know her states of mind. I will argue that the ethical pilgrim should not merely aspire to acquire reliable and accurate information regarding her mental states, she should also aspire to know her mental states in a way that allows her to shape them and, thereby, to unify herself as a speaker, thinker and doer.

I begin this argument by investigating two different ways to examine ourselves, first- and third-personal self-examination, that ground these two types of knowledge of our mental states. My investigation will be articulated as a response to the position put forth by Timothy Wilson, perhaps the most influential social psychologist writing about self-examination and self-knowledge. Wilson's work incorporates the results of a number of interesting experiments that are meant to highlight the fallibility of our capacity to examine ourselves. He takes these

to throw into question our capacity for introspection (or, as I will call it, "first-personal self-examination"). The remedy, he suggests, is to examine ourselves "from the outside" (or, as I will call it, with "third-personal self-examination").

I start the chapter with a short exposition of the evidence that leads Wilson to think that introspection is an unreliable form of self-examination. I will then introduce some conceptual distinctions that allow me to analyze Wilson's position with precision, disentangling distinct ideas that are often conflated in Wilson's arguments. We will come to see that Wilson's case against first-personal self-examination is significantly weaker than he takes it to be. Neither his theoretical considerations nor the empirical evidence on which he draws entitle him to hold that third-personal self-examination is more reliable than first-personal self-examination. What is more, Wilson himself recognizes that there are occasions when it is precisely the cultivation of our capacity for first-personal self-examination that improves our ability to know ourselves. I conclude the chapter by demonstrating that there is a kernel of truth in Wilson's position: some of the strategies by which we can make self-examination more reliable necessitate that we examine ourselves third-personally. In developing this claim I show, *contra* Wilson, that what makes third-personal self-examination more reliable than first-personal self-examination has to do with the kinds of motivated irrationality with which self-deception has been traditionally associated and not with the mechanisms that, according to Wilson, explain the unreliability of first-personal self-examination.

III.1 Timothy Wilson's Position

In 1977 Richard Nisbett and Timothy Wilson wrote a seminal paper in which they discussed different ways in which human beings tend to fail to know their higher mental processes through introspection.[1] A wealth of research followed this paper, much of which was gathered

1. Nisbett and Timothy D. Wilson 1977.

by Wilson himself, around 25 years later, into a comprehensive book on the topic: *Strangers to Ourselves: Discovering the Adaptive Unconscious*. The 2002 book, which addresses some of the criticisms that were raised against the 1977 article, has plenty to teach to psychologists, moral philosophers, and just about anyone interested in self-knowledge and its role in ethical development. In line with the article's spirit, Wilson's book highlights that introspection is more unreliable than one would initially imagine. Because of this, he suggests that it is often better to pursue methods of self-examination that are not introspective.[2] The closing paragraph of the introduction illustrates this:

> Socrates was only partly wrong that the "unexamined life is not worth living." The key is the kind of self-examination people perform, and the extent to which people attempt to know themselves solely by looking inward, versus looking outward at their own behavior and how others react to them."[3]

III.1.1 Explanations for Introspection's Failures

Wilson mentions two sets of explanations that account for our failures to know ourselves through introspection.[4] The first set of explanations is likely familiar to most readers. These explanations have been traditionally used to explain our failures to examine ourselves and have been exploited and put on display in many literary works. According to Wilson, within this traditional framework it is said that a person's introspection fails because:

1. the person is motivated to have an overly positive view of herself, making it hard for her to acknowledge mental states that are at odds with her own positive view about who she should be;[5]

2. Timothy D. Wilson 2002, pp. 16, 183.

3. Timothy D. Wilson 2002, p. 16.

4. Later in the chapter I will problematize Wilson's use of "introspection." For the time being, I will simply go along with his usage of the term.

5. Timothy D. Wilson 2002, p. 90.

2. the person is out of touch with her inner world and fails to identify the relevant issues that are moving her;[6]

3. the person is confused and cannot provide accurate reports about herself;[7]

4. the person is blinded by her hubris;[8]

5. the person represses some mental states because they bring about anxiety or psychic pain.[9]

Although Wilson concedes that these explanations account for some of introspection's failures, he does not spend much time discussing them. In his book he focuses, instead, on a second set of explanations which have emerged out of recent psychological research, explanations that invoke what Wilson calls the "adaptive unconscious," i.e. the module, or modules, of the mind which process information automatically and efficiently, but unintentionally and unconsciously. Wilson spends most of the book enlisting results from social and cognitive psychology that are meant to show that when subjects are asked to make reports about their minds they often end up confabulating stories based on faulty data. According to Wilson, when asked to report about themselves, subjects often end up:

1. relying on thoughts or feelings that were occurrent in their minds at that moment, even if these are not relevant or important to the issue being examined;[10]

2. identifying the aspects of the issue which are easiest to put into words, even if these are not the most significant;[11]

6. Timothy D. Wilson 2002, pp. 129, 160.

7. Timothy D. Wilson 2002, p. 4.

8. Timothy D. Wilson 2002, p. 3.

9. Timothy D. Wilson 2002, p. 6.

10. Timothy D. Wilson 2002, p. 171.

11. Timothy D. Wilson 2002, p. 171.

3. responding by appealing to shared cultural theories or idiosyncratic theories which describe how or why people in general (or this person in particular) typically respond, or should respond, in the way that they feel, think, or act, even if this is not an explanation for how or why they actually felt, thought or acted;[12]

4. relying on faulty observations of correlation or covariance between antecedent and subsequent events (correlations like: "Whenever I go to big parties I get depressed," or "I am in bad moods when I get less than seven hours of sleep").[13]

5. Reporting on phenomena that are, simply, inaccessible through introspection.

III.1.2 A Glance at Wilson's Evidence

Strangers to Ourselves is populated with references to experiments in social psychology and cognitive science that are meant to provide extensive support to his thesis that introspection is unreliable. A few examples will provide the reader with an idea of the kinds of empirical evidence that grounds Wilson's claims.

The Panty Hose Experiment

Subjects were asked to rate the quality of four pairs of panty hoses displayed on a table in a supermarket. Unbeknownst to them, the four pairs were identical. Most of the subjects, however, preferred those on the right of the display. There is plenty of empirical support showing that people have a marked preference for items on the right side of a display. When the subjects in Wilson's experiments were asked why they preferred the item they chose, they pointed to different attributes of their preferred item, such as its superior knit, sheerness, or elasticity. But no one mentioned that the position of the panty hose had anything to do

12. Timothy D. Wilson 2002, pp. 108, 171.

13. Timothy D. Wilson 2002, p. 108.

with its location. "When we asked people directly whether they thought that the position of the panty hose had influenced their choice, all participants but one looked at us suspiciously and said of course not."[14]

Post-Hypnotic Suggestion

Post-hypnotic suggestion leads subjects who have been hypnotized to do things with no conscious awareness of why they are doing them. If these subjects are asked why they are doing what they are doing, they usually come up with fabricated excuses, excuses that are utterly false but which the subject tends to believe nevertheless.[15]

Reflecting on Our Intimate Relationships

Wilson and other psychologists asked one group of college students in dating relationships to list the reasons why their relationship with a romantic partner was going the way it was and to then rate how satisfied they were with this relationship. Another group of students were merely asked to rate their satisfaction without any prior analysis of the reasons why the relationship was going the way it was. This latter group was meant to respond by appealing to their gut reactions. The ratings of the people in the "gut feeling" group predicted much better whether they would still be dating their partner several months later than those who analyzed the reasons. Wilson writes "[w]e have found that the feelings people report after analyzing reasons are often incorrect, in the sense that they lead to decisions that people later regret, do not predict their later behavior very well, and correspond poorly with the opinion of experts."[16]

14. Timothy D. Wilson 2002, p. 103.

15. Timothy D. Wilson 2002, p. 95.

16. Timothy D. Wilson 2002, p. 170.

Attraction on the Bridge

Two groups of males were interviewed by an attractive female. Some were interviewed on a scary flimsy footbridge that spanned a deep gorge while others were interviewed after they had crossed the bridge, resting on a park bench. When subjects completed the questionnaire the woman thanked them and said she would be happy to explain the study in more detail when she had time. The researchers kept track of how many of the males telephoned her later and asked her out on a date. Wilson mentions that, because the bridge felt dangerous, the subjects approached on the bridge were perspiring, short of breath, and had a rapidly beating heart. The researchers predicted that these men would be mixed up about exactly why they were physiologically aroused, attributing some of their arousal to being attracted by the woman. Wilson reports that "[t]his is exactly what seems to have happened. Sixty-five percent of the men approached on the bridge called the woman and asked for a date, whereas only 30 percent of the men approached on the bench called and asked for a date."[17]

According to Wilson, these experiments, and many others that I have not discussed, show that our capacity for self-examination is much more unreliable than what one would have initially thought. They are also supposed to bring out how pervasively we confabulate. Wilson also takes these experiments to call into question our capacity for first-personal self-examination *vis-à-vis* third-personal self-examination. I will return to Wilson's claims below. Before doing so, however, I will need to do some conceptual work and to establish some important distinctions that will allow us to better assess Wilson's position.

17. Timothy D. Wilson 2002, pp. 101-2, 130. The literature refers to this experiment as the "love on the bridge" experiment. The right attitude to attribute in this case is not love but attraction. I have altered the standard description to reflect this.

III.2 First- and Third-Personal Self-Examination

The aim of the previous section was to provide the reader with a flavor for the kinds of arguments and empirical evidence that underlie Wilson's claim that introspection is unreliable. In the next section I will develop some terminology that will allow me to delineate how I will understand the locutions "first-personal self-examination," "third-personal self-examination," "first-personal self-knowledge" and "third-personal self-knowledge." This terminology will help us assess Wilson's position.

Even though I will discuss this in more detail later (section III.2.5), it is worth anticipating that the way in which I will cash out these distinctions cuts across the different theories that have been offered by epistemologists to account for our first-person authority (that is to say, for what epistemologist usually refer to as self-knowledge). My account will not depend on the truth of any of these theories. And I will avoid wedding myself to any of them. This, I think, is one of the virtues of my account.

III.2.1 Looking Outwards

I have mentioned that one of Wilson's aims in the book is to argue that, because introspection is so unreliable, we should cultivate other methods of self-examination. Wilson often characterizes "introspection" with the expression "looking inward,"[18] and he contrasts this mode of self-examination with a different method of self-examination that he calls "looking outward." I will now proceed to analyze these two pairs of terms. I'll start with the latter.

Wilson devotes two whole chapters in the book to spell out what "looking outwards"

18. Timothy D. Wilson 2002, pp. 16, 69, 134,183,195,233,

amounts to and to highlight why it is valuable for self-examination.[19] According to Wilson, examining oneself by looking outwards includes learning about oneself:

1. by taking the perspective of an outside observer and inferring one's internal states from one's behavior,[20]

2. by participating on individual empirical tests that can assess one's mental states.[21]

3. by reading reports of controlled psychological studies,[22]

4. And by becoming acquainted with the way other people see one.[23]

According to Wilson, all these forms of self-examination fall under the umbrella of methods of self-examination that "look outward." Wilson does not explain what gives unity to all of these methods of self-examination. Arguably, all of them are unified by the fact that we use them to learn about ourselves by looking at ourselves from a distance, so to speak. In the last three methods of the above list, this distance is typically obtained by the mediation of an actual third person (the experimenter who performs the test, the researcher who writes the scientific report, or the friend or coworker who tells us how she sees us). Each of them can provide us, in one way or another, with knowledge about ourselves. However, what unifies all of these methods of self-examination that Wilson calls "looking outward" cannot be that there is an actual third person mediating our knowledge. The first method on the list advises us to take the perspective of an outside observer. But it is clear that the activity of observing our behavior and inferring our internal states from it can be performed by ourselves. This first method does not require a third person, it merely requires that we

19. Timothy D. Wilson 2002, Chapter 9 and 10.

20. Timothy D. Wilson 2002, pp. 204–211.

21. Timothy D. Wilson 2002, p. 194.

22. Timothy D. Wilson 2002, pp. 183–194.

23. Timothy D. Wilson 2002, pp. 194-203.

look at ourselves *as though* we were in the position of a third person. Once we recognize that a mediating person is not at play in the first method of self-examination, we can also come to see that it does not require it in most of the others. For instance, the person herself could be the only one responsible for setting up the experiment, recording the observations, analyzing the data and reaching the conclusion.

My proposal is that what unifies these methods of self-examination has to do with the fact that they are available, at least potentially, to a third person. These methods of self-examination are characterized by the fact that the person explicitly and self-consciously understands her examination to modeled on ways of examination available to a third-person. As such, they require that the person explicitly conceives her findings to be grounded on considerations to which a third person could likewise appeal. Because this third-personal perspective is central to these otherwise diverse forms of "looking outward," I will use the expression "third-personal self-examination" to refer to them.

The distinction that Wilson is drawing between these two methods to examine oneself is one that any theory of self-knowledge should be able to accommodate. Wilson is drawing on our intuitions about two different ways to examine ourselves, intuitions that are not meant to depend on the truth of any abstract epistemological theory about self-knowledge. I will discuss this in more detail below (section III.2.5)).

I prefer the locution "third-personal self-examination" over Wilson's "looking outward." To begin with, I think that it is confusing to speak of "looking *outward*" to describe an activity whose ultimate aim is to look at ourselves, i.e. to look inward. Although there is certainly something "outward" about the methods of self-examination that Wilson groups under the heading "looking outward," what is outward is the perspective from which one looks at oneself, not the direction in which one's gaze is directed. It would be better to call these methods "looking inward from the outside" and to contrast them with methods where we "look inward from the inside." The locution "third-personal self-examination" success-

fully captures both that the ultimate aim of one's investigations is oneself, and that this investigation is made from an external perspective. An additional reason why "looking outward" is inadequate is that not all the methods that Wilson groups under "looking outward" actually involve *looking*. In the list I provided above, behavioral observation, inference and testimony are all methods of third-personal self-examination. But it is potentially misleading to characterize drawing an inference or hearing the testimony of someone as a form of perception, as a way of "looking" outward. The locution "third-personal self-examination" captures, in a way that "looking outward" does not, what is common to all three ways of acquiring self-knowledge. Moreover, it helps to remind us that what is distinctive of all these methods of self-examination is that they involve reaching a conclusion about our mental states by appealing to grounds to which a *third person* could have appealed to reach it.

III.2.2 Looking Inwards

I have said that one of the central aims of Wilson's book is to show that introspection is a very unreliable form of self-examination. There are different ways in which one might cash out "introspection" and it will be important to get clear on what exactly Wilson is referring to.[24] Working through these different ways to characterize introspection will equip us with a better understanding of how "introspection" is used by Wilson (and by many other empirical psychologists who work on these issues). More importantly, this will allow us to isolate the particular way in which introspection has to be understood if it is meant to be contrasted with third personal forms of self-examination, that is to say, if it is meant to be a mode of

24. Wilson is sometimes equivocal with the usage of this term, equivocations which turn out to be quite pervasive in the work of social psychologists who deal with the sort of issues that Wilson discusses. See, for instance, Nisbett and Timothy D. Wilson 1977; Pronin 2009; Hansen and Pronin 2012.
Wilson occasionally uses "introspection" as a synonym of "self-examination" broadly speaking, not for a specific mode of self-examination that is not contrasted with third-personal self-examination (see, for example, Timothy D. Wilson 2002, pp. 162, 167). This usage does not happen too often and, as far as I have been able to tell, does not lead him to make arguments that depend on using this term equivocally. In fact, it is relatively easy to identify when he is using the term in this way because, when he does, he characterizes "introspection" positively. By contrast, when introspection is meant to be contrasted with third-personal self-examination, Wilson is critical of it, concerned with showing that it is unreliable.

examination that is *not* third-personal.

If one asks many of the empirical psychologists working on these issues to offer a paradigmatic image that epitomizes introspection, they are likely to ask us to picture ourselves when we "sit down in a comfortable chair, rest our chin in our hand, and take a moment for self-reflection."[25] This image is, of course, a caricature. Few of us ever examine ourselves in this way. Moreover, in empirical studies about "introspection," what subjects are asked to do is, more often than not, to fill in questionnaires. Thus, what the literature on empirical psychology seems to assume is that, if the subject is providing a self-report, then this self-report is the product of introspection. What is common between the image of the introspective subject in the couch and the person filling a questionnaire in the lab is that both are taking some time off their ordinary daily lives to reflect, by themselves, about themselves.

Wilson's thesis that introspection should be replaced by third-personal ways of self-examination depends on an understanding of introspection that characterizes it as a mode of self-examination that is *not* third personal. Introspection, understood as the self-reporting that takes place while sitting in a couch reflecting on oneself or filling questionnaires about oneself in a lab, is not a mode of self-examination that can be contrasted with third-personal self-examination. The person in the lab might be responding a questionnaire by self-consciously inferring conclusions from her observed external behaviors or by reflecting on an empirical study that she recently read. One can examine oneself third-personally both sitting in a couch or in a lab's chair, chin or pencil in hand. And this means that introspection, in the sense of an activity that is contrasted with third-personal self-examination, cannot be singled out by the fact that one is taking some time off to provide a self-report.

It is plausible to try to characterize a mode of self-examination that is contrasted with third-personal self-examination by appealing to a feature that, one might think, singles it out. It is common to think that something that goes along with an introspective examination is the

25. Hansen and Pronin 2012, p. 345.

speediness with which the person responds. When a person is responding to a question about her mental state through introspection, she typically responds readily.[26] By contrast, when you are examining yourself third-personally, you often need to spend some time collecting the evidence, organizing it, and assessing it. It usually takes some time to reach a conclusion in this way. "Readiness to respond," however, does not single out introspection as a mode of self-examination that is *not* third-personal. There are cases (and we will discuss a few of these later in the chapter) where a person is able to examine herself very quickly but in a third personal way. There are also cases where it can take quite some time to report on a mental state first personally, for instance because the person needs to put herself in a frame of mind that allows her to connect with an elusive emotion. I will discuss in more length a case like this below (section III.2.4).

As I have mentioned, Wilson often deploys the locution "looking inward" to refer to the mode of examination that is contrasted with third-personal methods of self-examination. There are moments where this contrast seems to be cashed out, both by Wilson and by other psychologists defending similar views,[27] in terms of observing two distinct types of *things*: a person's internal world, on one hand, and her outer behavior, on the other. This proposal lines up nicely with the vocabulary that Wilson deploys to describe these two kinds of self-examination; One form of examination looks inward, to the inner world of the person, the other looks outward, to her outward behavior.

This contrast between observing "two distinct types of entities" is predicated on a sharp separation between the inner and the outer, between desires, intentions and emotions on one hand, and outer behavior, audible speech and visible expressions, on the other. But by attempting to identify introspection with an examination of one of these two types of entities will not allow one to single out introspection as a form of examination that is *not*

26. See, for instance, Timothy D. Wilson 2002, pp. 168, 173; Hansen and Pronin 2012, p. 346

27. See, for instance, Timothy D. Wilson 2002, p. 183 or Hansen and Pronin 2012, pp. 346-9, 354.

third-personal. When a person is examining her overt behavior, she is not trying to look at an entity that is independent of her inner world. Rather, she is hoping to use this outward behavior to discover her internal states. The inner and the outer are not distinct and independent entities but rather two aspects of a single one: the mental life of a person. Outer behavior provides a different route to discover a person's inner world. And this means that when we understand introspection in this way, third-personal self-examination ends up being a species of introspection; "looking outward" is always a way to get at what is inward. Thus, if we understand introspection in this way we again fail to single out a form of examination that is *not* third-personal.

There is one proposal left, one that is, not only intuitive, but which Wilson seems to have in mind quite often when he talks about a mode of self-examination that is not third personal. Like the last one, this one is also anchored in the idea that introspection consists in "looking inward." Unlike the last one, this one is rooted, not in the location of what one is examining, but in the way in which one examines it. Throughout the book, Wilson suggests that introspection, when successful, provides us with direct access to our mind and direct knowledge of it.[28] Wilson often links the idea that this knowledge is *direct* with the idea that it is a form of knowledge which is thought to be *private* and *priviledged*.[29]

These last remarks might tempt one to characterize Wilson as an inner-sense theorist. Inner-sense theories postulate that we have an inner-sense, a (perhaps quasi) perceptual capacity, that provides us with direct access to our minds.[30] This perceptual capacity is meant to explain the immediacy and privacy of what seems to be our epistemically distinctive knowledge of our own minds. But this is a temptation that we should resist. Even if inner-sense theories were wrong, Wilson's proposals would still hold. Wilson is not committing

28. Timothy D. Wilson 2002, pp. vii, 7, 15-8, 23, 68, 73, 90, 108, 162-3, 183, 194, 205, 219.

29. Timothy D. Wilson 2002, pp. 17, 105, 108.

30. See, for example, Armstrong 1981; Lycan 1996.

himself to the view that we have access to our mental states through an inner sense.[31] The distinction between first- and third-personal forms of self-examination is meant to be intuitive, and to cut across epistemic theories of self-knowledge. Introspection, as Wilson is using it, is the mode of self-examination that is subjectively felt to be immediate and private. Because it is not a form of third-personal self-knowledge, it is a form of examination where the person is *not* trying to acquire information about herself by replicating the procedures of a third person trying to figure out her mental states. Introspection, as Wilson deploys it, is not a form of self-examination that is direct, private and privileged. If some epistemic theories of self-knowledge are right, we might not have private and privileged access to our minds. Introspection is, rather, the form of self-examination that *is taken to be* direct, private and priviledged, the form of examination that we intuitively think is exclusive to the subject and not available to anyone but the subject. (In the next chapter I will say a bit more of what warrants taken it to be like this). Thus, Wilson's idea that our access is "direct," in our context, should not be taken as a metaphysical claim about the nature of self-knowledge, but rather as a phenomenological feature that describes how we take ourselves to relate to this knowledge pre-theoretically.

Although there will be occasions later in the chapter where I will refer to this mode of examination as "introspection," for the most part I will prefer to refer to it as "first-personal self-examination." I prefer to use this locution to prevent the confusions that might arise from an equivocal use of "introspection" in the literature and to have a more consistent vocabulary.[32]

31. In fact, there are moments in the book where he appears to distance himself from this view (See, for instance, Timothy D. Wilson 2002, pp. 160, 173).

32. There will be occasions where I will want to leave open whether self-examination is first- or third-personal. When I do this, I will simply talk about "self-examination" without qualification.

III.2.3 These Two Methods Can Work in Tandem

These two forms of self-examination need not work in isolation. There are occasions when a person will come to know her mental states through a combination of them. There are plenty of cases where one can acquire self-knowledge from a combination of both first- and third-personal self-examination. Allow me to retool an example discussed by Finkelstein to illustrate this. A friend asks Hellen whether she loves her new boyfriend, Harry. She replies: "Well, I feel comfortable when I'm with him, and I am really attracted to him. But he talks too much about his therapy sessions and I don't always like his politics. Oh... also... you know how I can't stand being around people when they are sick? Well... When he had the flu I stopped every evening to bring him food and check up on him. So, yes, I suppose I love him."[33] In the case that I am imagining, Hellen is at a moment in her relationship where she is gradually coming to acknowledge that she loves Harry. Her feelings, however, are not yet fully fledged and she needs to complement what she can report based on how she feels first-personally with third-personal evidence. A case like this one illustrates that there are cases where a person comes to know her mental state through a combination of first and third person examination.

III.2.4 First-Personal Self-*Examination*?

The idea that introspection is a form of examination has an odd ring to it. Some readers might find it a bit strained to speak of such a thing as first-personal *"self-examination."* Think, for instance, of a case in which you ask a person about her desire for dessert. She will, most likely, respond immediately, off the bat, describing whether she wants dessert or not. She does not seem to be going around rummaging for clues, making inspections, or following steps that are meant to help her find out the answer to this question. No

33. I am borrowing this example from Finkelstein 2003, p. 123.

investigation appears to be taking place here. The person does not seem to be examining anything. She seems to be just reporting, off the bat, what is taking place within her.

I agree that in many of the cases where a subject makes a first personal report it can sound strained to describe this report as the product of an "examination." Nevertheless, I want to make the case that it is legitimate to speak in this way. To do so, I will describe a particular example where the person seeks to answer a question through first-personal self-examination, but where it is easier to see that calling it "examination" is appropriate. Having this case clear in sight will allow us to see how "examination" also gets a grip in the other cases.

Think about an occasion in which you are asked how you felt about someone you met in at party a few weeks back. At first, you might not really know what to say. You remember that you had a strong negative reaction, but you can't quite put your finger on it. So you try to think back to your experience in the party, placing yourself back in it. You try to remember the environment. Perhaps you hum some of the songs that were played, remember carefully the clothes you were wearing, think of the room where you spent most of the time, or attempt to visualize the portrait of the person about whom you are asked. As you do this, it suddenly dawns on you. Yes, you did not like this person; you found him petulant and arrogant. You can then respond to the question about what you felt for him. Here we have a process of vivid re-imagination that requires time, but which ultimately provides you with first-personal access to a mental state that was not readily available. I think that it is perfectly natural to call this process an "examination."[34] It takes time and requires you, as it were, to rummage and find what you are looking for. It is a process where you are genuinely looking for something and not merely reporting off the bat. What you are trying to find, though, is not, like in the case of third-personal self-examination, pieces of evidence from which you can draw a conclusion about what you felt. Instead, what you are looking

34. It might be interesting to mention that Wilson has some suggestions about how to improve our self-knowledge that go along these lines. (See Timothy D. Wilson 2002, pp. 156–8, 174–5)

for is a way to reconnect yourself with what you felt at the party, to reconnect with the feelings about which you are asked.

Having in view a case like this, where it is straightforward to say that the person is *examining* herself, allows us to revisit cases where it seems strained to use the notion of "examination." Think again of the case where a person is asked about her desire for dessert. The person will typically need to perform some sort of examination. Even if she is conscious of her desire for dessert, it is likely that she will not be actively attending to it. If so, she will need to redirect her attention to it. And when she does, she will recognize features of which she might have not been aware when she was not actively paying attention to it. Thus, even in this case, where the person is able to respond to a question about her mental states readily, it is still the case that there is a certain examination taking place. The person will need to connect with her desire for dessert, focus her attention on it, and experience it fully if she wants to produce an accurate report about such desire. Arguably, although this process might take place very quickly, it is still a process that makes sense to call "examination." First-personal self-examination is a form of examination because it requires that one focus one's attention on the relevant mental state in order to report on it. Thus, even in the cases where a person responds readily to a question about her mental state, the idea that she is examining herself does get a grip.

III.2.5 How These Distinctions Map onto Traditional Debates in Epistemology

I mentioned earlier that any theory of self-knowledge should be able to accommodate the distinction that Wilson is drawing between first- and third-personal self-examination. Some epistemologists might be asking themselves whether there is any way to account for this distinction within certain theories of self-knowledge, particularly within inferentialism. Inferentialism about self-knowledge is a family of deflationary theories that propose that there

is nothing epistemically distinctive about introspection. When I self-ascribe a mental state, even when I do it through introspection, I do so by performing the kind of inference from my behavior that an external observer could do to attribute this mental state to me.[35] Thus the inferential tradition portrays the authority that a subject has over her mental states as the authority of a reliable witness. The alleged special authority that the subject is meant to have over her own mental life comes from the fact that she has more information about herself than anyone else. From this perspective it would seem that it is one of the aims of inferentialism to do away with introspection. This is true when one is thinking of introspection as a mode of knowing that is ultimately non-inferential. But this is not how I am conceiving of introspection. I am not concerned with the ultimate grounds of introspection, but how introspection shows up in our ordinary lives. And I think that it is uncontroversial to say that introspection appears to us as non-inferential. Inferentialists can grant this. They can grant that when you introspect you are not taking yourself to be making any inference. This is compatible with thinking that this is an illusion and that, unbeknownst to you, your report is ultimately dependent on an inference. In fact, inferentialists have explained that the alleged privacy and authority of first-personal self-examination depends on the fact that these inferences occur effortlessly and automatically.[36] This automaticity and effortlessness is what gives subjects the impression that they are accessing their mind in an immediate way and not through an inference.

Wilson is not concerned with explaining the distinctive immediacy and authority that we have over our mental states. His use of "introspection" is not in the service of the metaphysics of the mind with which epistemologists are concerned. His project is concerned, rather, with understanding how human beings can know themselves better. When he invites us to examine ourselves in certain ways and not in other ways, it is irrelevant, for

35. Some defenders of interentialism are Ryle 1949; Byrne 2011; Carruthers 2011, 2010; Cassam 2014.

36. See, for instance, Carruthers 2010, p. 93.

his purposes, whether what he calls "introspection" is a process that is, ultimately, inferential. To see this, just imagine that inferentialism were true and all self-ascriptions were ultimately inferential. Wilson's invitation to avoid first-personal self-examination would still hold. He would say "You should avoid examining yourself through these methods that feel non-inferential (even if they are ultimately grounded on inferences). Instead, seek methods that are more intentionally and self-consciously modeled on the way in which other people know can come to know about your mental life."

According to inferentialists, there would not be a principled distinction between the ways in which you examine yourself first- and third-personally. But even if the is not principled distinction between these two ways to examine oneself within inferentialism, this theory can (and should) still have room to account for the distinction that Wilson is drawing between these two ways to examine ourselves. After all, there are cases where we can examine ourselves in ways that feel first-personal, ways that we feel are only available to us. And there are also cases where our investigation is self-consciously pursued like the investigation that a third-person could in principle pursue about our mental states.

The fact that this distinction is not a principled one for inferentialism, however, generalizes to other theories of self-knowledge. Take inner-sense theories. These theories, I mentioned, postulate that what makes our knowledge of our mental states special is that we have a (perhaps quasi) perceptual capacity that affords us a direct and (perhaps) priviledged access to our own minds. If you are an inner-sense theorist, it might be tempting to identify first-personal self-examination with the exercise of inner-sense. This identification, however, might not be correct. Even if inferentialists are wrong about the nature of self-knowledge in general, they are right that there are occasions where we end up making claims about ourselves in inferential ways that we fail to acknowledge as inferential (this is one of the positive conclusions of this chapter). Wilson is not committing himself to the view that we

have access to our mental states through an inner sense.[37] His point is neutral with respect to the epistemological theories of self-knowledge. When our examination is grounded on inferential processes that we don't acknowledge, when we attempt to report on the contents on our mind in the way that we intuitively feel to be private and direct, this examination is first-personal. Because in any theory of self-knowledge that is not inferentialist, there will be cases of self-examination that feel immediate and direct but are actually inferential, no theory of self-knowledge will be able to make a principled distinction between first- and third-personal self-examination.

III.3 Wilson and the Alleged Limits of First-Personal Self-Examination

In their seminal paper about self-examination and self-knowledge, Nisbett and Wilson argued that people's self-examination is very unreliable. They discussed a number of experiments that were meant to show that people often misreport the causes of their behavior, either by falsely reporting influences that were not operative or by failing to acknowledge influences that were operative. Their conclusion was that people are rarely more accurate in explaining their own behavior than outside observers who were guessing based on information about the public features of the situation and the subjects' responses.[38] This paper has been one of the most cited articles in social psychology. When it is cited, it is often cited as the source of an authoritative truth that was well established almost forty years ago.

Of course, Nisbett and Wilson's article has also been criticized.[39] In the 2002 book Wil-

37. As I mentioned, there are there are moments in the book where he seems to disavow this position (Timothy D. Wilson 2002, p. 160).

38. Nisbett and Timothy D. Wilson 1977.

39. For instance, the article relies on the distinction between the content of a mental state and the processes that brings this content about. In their study Nisbett and Wilson grant that subjects have introspective

son revised his earlier position and conceded some ground to these criticisms. He reworked the distinction between process and content and granted that introspection is not as unreliable as he and Nisbett had formerly suggested. His concessions, as I will argue, do not go far enough. The terminological work done in the previous section will allow me to undertake a careful scrutiny of Wilson's views. Here are some of the main conclusions that I will argue for. I will first show that Wilson's justifications for the claim that first-personal self-examination is unreliable are inadequate, from both a theoretical and an empirical perspective; the "adaptive unconscious" is not as inaccessible as he suggests, and the experiments do not properly warrant his conclusions. I will also argue that Wilson's empirical evidence fails to control for the key variables and that some of his concrete recommendations actually run counter to his own thesis. Finally, I will argue that, even though it might be true that there are many cases where first-personal self-examination is less reliable than third-personal self-examination, the explanation for why this is the case is not the explanation that Wilson provides.

III.3.1 How Much Is Introspectable? Wilson's Theoretical Justifications

As I was saying, in the 2002 book Wilson attempts to address some of the main challenges that were raised against the views he had defended in the articles co-authored with Nisbett. He reworks the earlier distinction between content and process in terms of a distinction between mental states caused either by the 'conscious self' or by 'the adaptive unconscious.' He now grants that "to the extent that people's responses are caused by the conscious self,

---

access to the content of their mental states, but deny the that they have introspective access to the processes. This distinction has been criticized on the grounds that it is impossible to properly draw such a distinction. As one of these scholars put it: "there are no criteria by which a mental event could be located under one heading and discriminated from the other" (White 1980, pp. 105–6). For other criticims along this line, see: E. R. Smith and Miller 1978; White 1988, 1989.

Scholars have also argued, and in this they side with common sense, that it is implausible to think that we do not have introspective access to our mental states and that agents are are often better at knowing themselves than observers. (For a list of references to studies that establish this latter conclusions, see Newell and David R. Shanks 2014, p. 5).

they have priviledged access to the actual causes of these responses."[40] But, despite this concession, he still holds that "*much* of what we want to know about ourselves resides outside of conscious awareness."[41] This is so because he holds that the majority of our behavior is caused by the 'adaptive unconscious.'

Because he believes that conscious awareness is a prerequisite for self-knowledge that arises out of first-personal self-examination, and because he identifies conscious awareness with the activities of the conscious self, this leads him to assert that we don't have first-personal access to much of what we want to know about ourselves. His theoretical arguments defending this view, however, are inadequate.

Wilson's thesis is not merely a statement of the truism that a good deal of what we could know about ourselves cannot be known through first-personal self-examination. We cannot examine first-personally the firing of our neurons, the chemical reactions in our brains or the electrical processes in our nerves. Nobody would deny this. What makes Wilson's thesis interesting is the claim that many of the things that have been traditionally thought to be first-personally accessible, things like beliefs, emotions, intentions or desires, are not as easily accessible as they have been thought to be. However, when Wilson defends the *inaccessibility* of our higher order mental states, his argument often relies on the inaccessibility of these sub-personal processes. Sliding from the register of higher order mental processes to the register of sub-personal processes is not a legitimate way to argue for this point.

There are other times when Wilson argues that if a mental state has been inferred from the observation of a person's behavior, and this mental state conflicts with the person's explicit self-report, then the person has no first-personal access to it. This view is inadequate in two different ways. First, and as Jennifer Nagel points out, "the fact that an attitude has been revealed by an indirect measure does not entail that it is unavailable for direct

40. Timothy D. Wilson 2002, p. 106.

41. Timothy D. Wilson 2002, 4. My emphasis.

discovery."[42] In some cases the self-reports are unreliable because subjects have incentives to present themselves in a particular way. The way in which the self-report is requested also makes a difference to the output. People who are asked to predict their implicit associations (rather than endorse them) can do quite well in providing reports that align with the indirect measures drawn from their outward behavior. Nagel mentions that if people are asked about their 'gut reactions' as opposed to their 'actual feelings', their responses are a closer fit to their responses on the Implicit Association Test.[43]

Second, Wilson fails to identify the proper modality of the claim: "much of what we want to know about ourselves resides outside of conscious awareness." Wilson thinks that if a mental state takes place beyond conscious awareness, then it is not accessible through first-personal self-examination. But this does not seem right. In so far as an unconscious mental state could, at some point, be made conscious, it is a mental state to which we *can* have first-personal access, regardless of the fact that we might lack such access here and now. Wilson might be right that a very significant part of our mental life takes place, not merely outside of conscious awareness, but in such a way that it cannot even be brought at will into conscious awareness here and now. But this does not mean that the relevant mental states cannot eventually be known first-personally.[44]

In the book Wilson suggests that the main reason why first-personal self-examination is unreliable is, quite simply, that it cannot be deployed to know much of what we want to know about ourselves. (This was the fifth item in Wilson's list of explanations that accounted for why introspection was unreliable that we discussed in section III.1.1). What I have said in this subsection, however, shows that his claim isn't adequately justified.

42. J. Nagel 2014, p. 236.

43. For further criticisms along these lines see Gawronski, Hofmann, and Wilbur 2006; Gawronski and Bodenhausen 2012; Newell and David R. Shanks 2014

44. My point, to put it in Aristotelian terminology, is that a higher order mental state is potentially conscious even if it has not achieved first actuality.

III.3.2 How Much Is Introspectable? Wilson's Empirical Evidence

The empirical results which are meant to provide support to the thesis that self-examination in general and first-personal self-examination in particular are unreliable are also problematic. There are a few problems with this. First, the ecological validity of the experiments is dubious. Second, Wilson has not given us proper grounds to argue that introspection is unreliable across the board. Third, the experiments do not provide a strong case for Wilson's thesis; there are compelling explanations for their results that do not involve failures of self-knowledge. Finally, Wilson fails to control for the key variable in his experiments. The first two objections have to do with self-examination in general. The latter with the specific claim that first-personal self-examination is unreliable.

Ecological Validity

In Wilson's work, there are two groups of studies whose ecological validity can be criticized along similar lines. The first has to do with studies, like the panty hose experiment, that by design fail to make transparent to the participants the relevant (and often contrived) features of such studies. In the case of the panty hose experiment, the participants were led to believe that the products on display were different. The experiment was meant to show that human beings are quite bad at identifying the reasons why they choose one product over another. But given the fact that the circumstances are so contrived, it is unclear whether this claim generalizes to ordinary circumstances. It is plausible to think that in real life scenarios, where participants are faced with four distinct products, the actual quality of the products will play a more significant role in determining their choices than the position of these products. A study like the panty hose experiment is, in some ways, like a study of perception which asks participants in a dark room to identify different colors and then uses participants' failures as evidence of their incapacity to identify colors in proper lighting

conditions.

Appeals to post-hypnotic suggestion can be grouped with a second group of studies of questionable ecological validity. Post-hypnotic suggestion provides a clear illustration of our capacity to offer *ex-post-facto* rationalizations that we do not acknowledge as such. Studies that depend on hypnosis can be grouped with a number of other studies that rely on the study of subjects in altered or pathological mental conditions (such as Michael Gazzaniga's split-brain patients[45]). Once again, the fact that these experiments rely on situations that are not standard (in this case, on the fact that subjects have been subjected to hypnosis or that their brain functioning is impaired) puts into question whether they can be generalized to standard cases. As Peter Carruthers has argued, arguments based on these kinds of examples resemble, in some ways, the argument from illusion in the philosophy of perception that is meant to show that we do not have direct perceptual contact with external objects. "The existence of illusions deriving from clever deceptions and brain manipulations doesn't show that I lack direct perceptual contact with my coffee cup in the normal case (at least not without considerable further argument). Likewise the occurrence of confabulation deriving from clever deceptions and brain manipulations doesn't, without considerable further argument, show that I am not in direct introspective contact with my judgments and decisions in the normal case."[46]

Is First-Personal Self-Examination Unreliable Across the Board

We can also doubt the ecological validity of these experiments from the perspective of our ordinary convictions about self-knowledge. It is extremely counter-intuitive to think that introspection is unreliable across the board or that an outside observer would be as reliable as us in speaking about our own mental states by making guesses based on information about

45. Gazzaniga 1995, 2000

46. Carruthers 2010, pp. 92–3.

the public features of the situation and our explicit responses. There are all kind of mental states that appear to be readily and correctly articulable by the subject and over which the subject seems to have special authority. As David Finkelstein points out in the opening of his book: "If you want to know what I think, feel, imagine or intend, I am a good person—indeed, usually the best person—to ask."[47] Saying this, of course, does not amount to the suggestion that the subject is incorrigible about her self-ascriptions. But it does suggest, as Victoria McGeer argues, that "the evidence of someone's 'unextorted word' about what they think, desire, or feel takes a lot to defeat. For not only must the rest of the person's behaviour speak strongly against taking them at their word; there must be some reasonable account of how they have failed to maintain first-person authority in the particular case."[48]

Acquiring knowledge of a person's mind through third-personal ways of investigation is often very difficult, and this difficulty suggests that third-personal self-examination will be unreliable. Wilson has done some experiments to show that people are not any better than strangers to predict the causal influences on their moods.[49] But this is not sufficient evidence to argue that, across the board, third-personal self-examination is as reliable as first-personal self-examination. Think, for instance, about the difficulty involved in figuring out some of my intentions. There is a lot that one has to know about my current and past behavior to be able to know that I am planning to stand up and go to the kitchen in the next two minutes. In fact, judging by my past behavior, an observer will likely judge that I am walking to the kitchen to grab breakfast. But the observer would be mistaken because today I want to get some more work done before eating; I only intend to grab a cup of water. Alternatively, think of the case where you are confused about a sentence in this chapter. The easiest way to figure out what I meant to say is to ask me. It would often be hard for a witness, particularly one that is merely making guesses based on few pieces of information, to figure this out.

47. Finkelstein 2003, pp. 1, 100.

48. McGeer 2007, p. 81.

49. Timothy D. Wilson 2002, pp. 110–3.

These are just two run-of-the mill examples that remind us of "natural common assumption that human beings have a special kind of authority with respect to claims they make about their own minds, in particular about their own intentional attitudes."[50] Of course, perhaps scientists will develop an Inner Sense Detector that is more reliable than our first-personal self-examinations. This is not a possibility that my remarks refute. What I am refuting is that the considerations to which Wilson appeals are sufficient to claim that first-personal self-examination is less reliable than third-personal self-examination across the board.

Psychologists and empirically minded philosophers might not be impressed by appeals to common sense or to what appear to be intuitive examples like the ones I just offered. Psychologists, however, have also brought up plenty of evidence that puts pressure on Wilson's counterintuitive proposal. Ben Newel and David Shanks, criticizing Nisbett and Wilson, refer to empirical research which shows predictive advantages for actors over observers.[51] This research suggests that the subject of the mental states has better self-knowledge than a third person observing her: "[i]t is apparent that in many of the sorts of situations cited by Nisbett and Wilson, we do in fact have introspective access to our conscious mental states, and the verbal reporting of these states conveys privileged information about the causes of our behavior."[52]

The design of Wilson's own experiments, however, often undermine his own claims. The experiment which was meant to show that strangers were as accurate in predicting the causal determinants of people's moods relied on correlating certain features of subjects' lives with their moods. These correlations, however, depended on subjects' own introspective reports on issues like their own moods and the quality of their relationships with friends, reports that are assumed to be accurate and, consequently, which undermine the ambitious version of the view that he is trying to defend.

50. McGeer 2007, p. 81.

51. Newell and David R. Shanks 2014, p. 5.

52. Newell and David R. Shanks 2014, p. 5.

In the next chapter I will argue that it is hard to make sense of our life as guided only by knowledge to which we are related in a third-personal way; third-personal self-knowledge has to be an exceptional way to know our mental states, not the norm. Wilson's experiments concerning self-knowledge do not deal with cases where we appear to have readily accessible and accurate knowledge of our own mental states but with cases where self-knowledge is difficult to attain. It follows from what I have been saying that we should qualify Wilson's suggestion that first-personal self-knowledge is less reliable third-personal self-examination. Instead of reading "introspection is less unreliable than third-personal self-examination" we should read "introspection is less reliable than third-personal self-examination *in certain cases where self-knowledge is difficult to attain*."

Multiple Levels of Causality

Even if we accepted the ecological validity of some of these studies, and even if we qualified Wilson's remarks to cases where self-knowledge is difficult to attain, his interpretations of the experiments still pose a further problem which Wilson does not discuss or address, despite the fact that some scholars have pointed it out.[53] Studies like the panty hose experiment or the attraction on the bridge experiment show that participants make a certain mistake. But these experiments do not unequivocally show that the mistake has to do with participants' self-knowledge.

Take the party hose experiment. Wilson claims that the participants in the panty hose experiment failed to identify the panty hoses' position as the factor that lead them to prefer the ones in the right of the display. He claims that this shows that they did not know the reasons for their action. At a certain level of description this might be true. Participants failed to recognize that the position of the products was the most important causal factor

53. See, for instance, E. R. Smith and Miller 1978

that explained their choices. However, given the claim that Wilson is trying to make, what is relevant is not merely whether the person can identify all the causal factors for their judgment about the quality of the panty hose regardless of their level of description. One plausible explanation for what happened in this experiment is that the position of the panty hose led participants to misjudge which one was better. But a participant who chose the fourth panty hose might have truly believed that its material was more flexible. And this (mistaken) judgment about the panty hose might have led her to assert that the fourth panty hose was better. If this is the case, then it would be true that the reason why this person chose this panty hose as the best one was that she thought it was more flexible. If so, this person knew the reason why she was making her decision.[54]

Something similar can be said about the attraction on the bridge experiment. It is possible that the heightened danger might have led to a change in the person's attention or to an increase in his arousal. These, in turn, would have caused the man to find the woman in the bridge especially attractive. If this were the case, then this would also be a case where the mistake is made in the participant's judgment about the woman in the bridge, not in his capacity to know himself.

The position of the panty hoses and the effects of the heightened danger influenced the participants. But what Wilson would need to show, if he does not mean to reduce his argument to triviality, is not merely that there was something causing participant's behavior that they did not know. What he would need to show is that what participants took to be causing their behavior was not, actually, causing it. As Eliot Smith and Frederick Miller have argued: "It certainly can be argued that the proximate cause of any action is a pattern of neurons firing. Yet, no subject can be expected to report on which neurons are firing, even though another experimenter with appropriate instruments might have access to that level of causation. Since we assume that Nisbett and Wilson do not mean to reduce their

54. Carruthers 2010 criticizes these experiments along the lines I have just sketched.

argument to triviality by identifying 'mental process' with neurons firing, it follows that this intrinsic ambiguity in the term 'cause of behavior' bars the use of a disagreement between subjects and experimenter as to what is the cause in some particular instance as evidence against introspective access to mental process."[55]

Thus, even if the experiments do bring out problems with our capacity to track important influences on our behavior, they are not conclusive enough to prove that we lack self-knowledge of our higher mental processes.

III.3.3 Wilson's Experiments Fail to Control for the Key Variable

The issues I have raised put pressure on the studies on which Wilson relies to make the case that we are not very good at self-examination in general. I'd like to criticize now some of the evidence that is supposed to ground Wilson's thesis that *first-personal* self-examination is less reliable than third-personal self-examination.

Wilson appeals to a number of studies to show that introspection is very unreliable. As I've argued, for his claim to be valid, "introspection" cannot be simply equated with "self-report," it has to be understood in a more restricted way, i.e. as "self-report that emerges out of first-personal self-examination." Getting clear on this brings out a very important flaw in Wilson's evidence. In none of the experiments that Wilson uses to make his case the experimenters are actually controlling for the key variable.

Let me illustrate this with two examples. In the attraction-on-the-bridge experiment. We are not told anything about the type of mental processes that led participants to call the interviewer. It might be natural to assume that the subjects concluded that they were attracted to the interviewer by first-personal examination. But this is not something that Wilson can assume if he wants to use these experiment to substantiate the empirical claims that he wants to put forward. Similarly, in the experiment where participants were asked

55. E. R. Smith and Miller 1978, p. 357.

to list the reasons why their relationship with their romantic partner was going the way it was, researchers did not give these subjects any instructions about the particular way in which they should access these reasons, nor did they design any mechanism that guaranteed that the reasons that these subjects provided were first-personal; they did not even ask these subjects to inform the researchers whether they were attempting to access these reasons first-personally. Given that Wilson is appealing to this experiment to prove that first-personal self-examination is unreliable, failing to control for this, failing to make sure that subjects were actually examining themselves first-personally, is a very significant problem in his evidence.

As far as I can tell, Wilson's failure to control for this key variable impugns the scientific rigor of his argument. I think that it is reasonable to assume that the subjects of the experiments that he mentions were accessing their mental states first-personally, however, the assumption isn't air-tight enough to justify his claim..

But it is worth mentioning that, even if these experiments show that first-personal self-examination is less reliable than what we might have first thought, it would not warrant the conclusion that third-personal self-examination is more or less reliable. This is not something that he can infer from the unreliability of first-personal self-examination provided by these experiments. These experiments might show that our self-examination is unreliable because we examine ourselves first-personally, but this tells us nothing about the extent to which third-personal self-examination is more reliable than first-personal self-examination. In fact, and as I will now argue, given the type of arguments that Wilson wants to put forth, it is not clear that he has any way to support this thesis.

The fact that these experiments are not controlling for this key variable is a significant problem for Wilson's scientific rigor. This might make some readers restless. Is it really possible that Wilson is not controlling for the key issue which he wants to demonstrate? And, if he does not, why should we take the time to examine carefully his flawed position?

The response to the first question is that Wilson is, in fact, not controlling for the key

issue. This important oversight is the result of his lacking the sufficient conceptual clarity that would allow him to recognize that this is, in fact, the key issue that needs to be controlled. Wilson does not work with the concept of "first-personal self-examination" but rather with "introspection." And he uses this concept equivocally, leading him to end up conflating different senses of the term. The end result is that the experiments always end up supporting the claim: "introspection is a fallible capacity," where "introspection" ends up being a moving target whose meaning alters depending on the case being discussed.

This leads us to the second question: "given Wilson's conceptual sloppiness, why take the time to examine his position?" My answer is two fold. First of all, the main thesis of Wilson's book has been quite influential in the literature on empirical psychology, a literature whose influence over philosophy is growing. It is relevant to examine Wilson's position because, in doing so, we are shedding clarity upon a burgeoning literature. My reason for dealing with Wilson's position, however, is not merely that it is becoming popular. The equivocations in the use of "introspection" are pervasive in the literature because they are quite natural. These equivocations are likely to haunt anyone who is interested in the role that self-knowledge plays in ethics. The family resemblances between different cases that can be called "introspection," the family resemblances among different senses of "looking inward," make it hard to see the potential equivocations involved in the different uses of this term. Examining Wilson's position allows us to get clearer on this.

My suggestion, in other words, is that it is worth spending our time examining Wilson's position because it is not at all obvious that it is flawed. Recognizing the problems in it requires the type of philosophical heavy lifting that I have been doing in this chapter. Addressing his position provides us with an opportunity to sharpen concepts that will help us to understand more clearly the nature of self-knowledge and its role in ethical development.

III.3.4 Wilson's Sympathies for First-Personal Self-Examination

Wilson asserts that we can remedy introspection's unreliability by examining ourselves third-personally. However, and as I will now argue, if we attend carefully to what Wilson says about some of the experiments, we come to see that his suggestions reveal that he is much more sympathetic to first-personal self-examination that what his rhetoric would have made us think.

In the book, Wilson does not offer suggestions about how to correct our tendency to confabulate in cases of post-hypnotic suggestion or cases like the attraction-on-the-bridge experiment. He does, however, suggest strategies to correct for our fallibility to report accurately about how we feel. Asking participants to articulate their reasons why their relationship with a romantic partner was going the way it was tended to be incorrect.[56] He suggests that the remedy to deal with these failures of introspection consisted in learning to develop informed gut feelings without analyzing these feelings too much.[57] Wilson suggests that, at least in many cases, what we should learn to do is to trust our feelings and go with them.[58]

Wilson is here instructing us *not* to analyze our feelings, not to infer them from our outer behavior or from generic theories about how we act or should act. Instead he invites us to trust our gut feelings and go with them. This proposals amount to a suggestion to learn to examine and know ourselves first-personally in this particular way. And this means, contra Wilson's own rhetoric, that his advice to avoid some of the difficulties involved in self-examination consist, precisely, in developing our capacity for first-personal self-examination (at least of a certain kind).

56. Timothy D. Wilson 2002, p. 170.

57. Timothy D. Wilson 2002, p. 172.

58. Timothy D. Wilson 2002, p. 172.

III.3.5 Does the 'Adaptive Unconscious' Affect First-Personal Self-Examination More Than Third-Personal Self-Examination?

Wilson's book is ripe with examples that purport to show that much of our higher order cognitive processes depend on mechanisms occurring outside of our conscious awareness. His book intends to demonstrate that our lives are determined, quite significantly, by unconscious processes that shape how we perceive and interpret the world, unconscious processes that dispose us to think and act in quite determined ways. I will now argue that it is quite natural to conclude, from Wilson's description of the functioning of the mind, that the exact same mechanisms that derail first-personal self-examination will also derail third-personal methods of self-examination. If this is so, then it will not be possible for him to argue, on these grounds, that third-personal self-examination is more reliable than first-personal self-examination.

Wilson appeals to a series of unconscious mechanisms to explain why first-personal self-examination often gets derailed. But it is natural to think that these unconscious mechanisms are operative in many different types of mental processes, not merely when these processes are concerned with examining ourselves first-personally. Wilson does not offer any principled reason to think that the mechanisms that affect first-personal self-examination will not be operative, also, when we examine ourselves third-personally. In fact, he occasionally brings out that these mechanisms derail self-examination in places of the book where he is speaking about self-examination in general, not about first-personal self-examination in particular.[59] In section III.1.1 (pp. 91) I offered a list of explanations that, according to Wilson, account for the mechanisms that explain how first-personal self-examination gets derailed. This list can be replicated, item-by-item, to explain why third-personal examination can go wrong. When

59. See, for instance, Timothy D. Wilson 2002, pp. 163-4

I examine myself third-personally it is likely that I will end up: 1) relying on current issues that are salient to my attention even if they are not relevant; 2) reporting on the aspects of the issue which are easiest to put into words; 3) responding by appealing to theories which describe typical behavior of people (or myself), theories which are not relevant to what I need to report; 4) relying on faulty observations of correlation between antecedent and subsequent events.

In the book Wilson focuses on certain explanations that account for why first-personal self-examination is unreliable, namely, explanations that appeal to the unconscious mechanisms at play the "adaptive unconscious." Because these unconscious mechanisms are also operative in third-personal ways of self-examination, Wilson is not entitled to claim, based on these explanations, that third personal self-examination is more reliable than first-personal self-examination.

III.4 A Defense of a Weak Version of Wilson's Thesis

III.4.1 Objective Self-Examination

Here are a few of the conclusions that we have reached thus far:

1. Although Wilson wants to argue that self-examination in general, and first-personal self-examination in particular, are unreliable, the experiments on which his argument relies are far from conclusive. Although these experiments bring out unexpected failures in the performance of participants, there are many ways to interpret these failures such that the failure is not a failure of self-knowledge but a failure of judgment. In most of these cases, this latter explanation is actually much more plausible.

2. It is not true that third-personal self-examination is more reliable than first-personal self-examination. In fact, Wilson himself provides arguments that helps us see that there are at least some cases where first-personal self-examination is a capacity that should be cultivated and developed because it helps to know ourselves (section III.3.4).

3. The types of explanations that Wilson deploys do not entitle him to argue that third-personal self-examination is more reliable than first-personal self-examination. Third-personal methods of self-examination are corrupted by the same mechanisms that, according to Wilson, make first-personal self-examination unreliable (section III.3.5).

Despite this, I do believe that there are cases, and they are quite pervasive in the case of ethical development, where third-personal self-examination might be more reliable than first-personal self-examination. My justification for this will be twofold. First, there are strategies to examine ourselves which can enhance the reliability of our self-examination. But these strategies can only be deployed in the case where we examine ourselves third-personally. From this it follows that there are certain types of third-personal self-examination that can be more accurate than first-personal self-examination. Second, I will bring out that the nature of first-personal self-examination is such that there is an ineliminable dimension of dogmatism built into it. This explains why it might be more difficult for us to identify cases where first-personal self-examination can be corrupted.

Wilson's book helps us to see that unconscious mechanisms are a significant source of what derails our capacity to examine ourselves accurately. We can secure more reliable ways to examine ourselves by devising strategies where our unconscious mechanisms are kept at bay. One way to do this is by outsourcing our examination to someone else. This is, of course, something that one can only do when one is examining oneself third-personally. You cannot outsource a first-personal self-examination.

When I described the nature of third-personal examination (sec. III.2.1), I mentioned that it was natural for readers to be tempted to think of it as requiring the actual mediation

of a third person. I mentioned then that this interpretation squares nicely with most of the methods of third-personal self-examination that Wilson discusses in the book: the person who learns about herself by reading reports on controlled psychological experiments does not normally take any part in the execution and analysis of these experiments; the person who takes individual psychological tests to know about herself is not typically involved in the process of collecting information or drawing conclusions from it; finally, the person who learns about herself from someone else requires the mediation of a third person. To the extent that the person seeking self-knowledge is kept out of the investigation, to that extent are her ordinary unconscious mental states prevented from interfering with it.

Suggesting that outsourcing our self-examination makes it more reliable is in line with what the natural and social sciences have recognized for many years now, namely, that blind and double-blind tests and experiments are more reliable than open trials. Devising concealing techniques that withhold from subjects, researchers, technicians, and funders, key information about the test or experiment is a way to keep their unconscious mechanisms from interfering with it.[60]

Although these types of results do not entitle us to conclude that third-personal self-examination is more reliable than first-personal self-examination across the board, it does suggest that *certain* forms of third-personal self-examination can be more reliable than first-personal self-examination. It is important to note, however, that although it is natural to think that outsourcing the processes of observation and analysis to a third person improves the investigations' reliability, this need not always be the case. What is crucial for the examination to be more reliable is that the mental mechanisms that interfere with the investigation are kept at bay, not that a third person practices them.

60. It might be worth highlighting that what I am saying here does not only apply to experiments which are concerned with knowing ourselves. As I mentioned earlier, the unconscious process that, according to Wilson, interfere with self-examination, are operative simply when we exercise our minds. As such, they are operative in any investigation in which we engage, not just in investigations about ourselves.

To begin with, the examination will not be more reliable if the third person to whom we outsource the examination is also misled by mental processes like the ones that mislead our own attempts to examine ourselves. This could happen, for instance, because our own self-reports get in the way of this person's supposed independent assessment or because this person happens to share our same biases, predispositions or preferences. When Wilson invites us to listen to what others think about us, it is no use to listen to what they think about us given what we ourselves have told them. Similarly, the results of the experiments done by a third-person are not likely to be free of our own biases if we take an active part overseeing or supervising them.

The fact that we outsource an investigation need not make the investigation more reliable. Outsourcing our investigations to someone who is more biased than we are, or to someone whose mental mechanisms are more likely to interfere with the investigation is worse than pursuing these investigations on our own. Of course, many of the biases that get in the way of our self-examination have to do with issues related with self-worth and our own self-image. Because our self-worth and self-image are not usually prominent in a third person, it is often valuable to rely on them to help us conduct the examination. But, once again, this will work only if the people we ask can detach emotionally from the investigation. People who are very close to us, like our immediate family members, are quite likely to be as liable to be misguided as we are.

Wilson suggests that we can know ourselves better if we tap into what other people think about us. He never problematizes the fact that, for this advice to work, who you ask is important. In a recent article that follows closely Wilson's footsteps, the authors are more careful: "[Self-knowledge] can sometimes best be obtained by looking outward to the opinions of others, *especially those wiser than ourselves and with more detached perspectives.*"[61] The qualification to the assertion is very significant. What is crucial when you outsource your

61. Hansen and Pronin 2012, 346, my emphasis.

self-examination is not that you outsource it to a third person, but that this person is *wiser* and *more detached* than you.[62]

To prevent some of these unconscious mechanisms from interfering in our self-examination, it is not necessary to actually appeal to a mediating person. It is possible, at least in principle, to devise methods of observation and analysis which the person herself can carry out judiciously and which can prevent the intrusion of these derailing mechanisms. There are many tests where a computer is responsible for recording the information, organizing it and reaching some kind of verdict (think, for instance, of an Implicit Association Test). If the information is easy and straightforward to record, and the algorithm to organize these observations and reach a verdict simple to follow, then the person herself could industriously follow the computer's algorithms and reach the same result. In such a case there would be no third person involved; the person who is examining herself would have been the single agent responsible for the observation and analysis of the information. But it is perfectly plausible to think that, if the protocols are straightforward and simple, following them can be successful in precluding the interference of the person's own disruptive unconscious mental processes. Self-examination is more reliable, not so much when it is pursued by a third person, but when there are protocols that ensure that the disruptive mental processes of the investigator do not get in the way of the examination.

This suggests that what makes certain ways of third-personal examination more reliable is not so much that they are performed by a third person but that they are performed in what I will call, for lack of a better expression, an "objective way." The locution "objective" is mainly meant to indicate that the examination is pursued in such a way that the subject has been taken out of the way so that her unconscious mechanisms do not interfere. We should keep in mind, however, that these protocols are meant to preclude these interferences

62. Unfortunately the authors of this article, who had written this qualification in their introductory comments, failed to keep it in mind later in the paper, when they actually discussed how the opinion of other people about us could improve our self-knowledge (Hansen and Pronin 2012, pp. 354–7).

in any person, not just in the subject who is being examined.[63]

III.4.2 The Adaptive Unconscious and the Freudian Unconscious

The actual mediation of a third person will only contribute to prevent *some* of the unconscious mechanisms that derail self-examination. I mentioned earlier that, in the book, Wilson is mostly interested in discussing certain types of unconscious mechanisms that interfere with our self-examination. According to him, these unconscious mechanisms do not result, like the unconscious mechanisms discussed by Freud, from hidden motivations; they are the result of the particular workings of our minds that Wilson calls the "adaptive unconscious."[64] I offered a list of these mechanisms at the beginning of the chapter (section III.1.1). If we review this list we come to realize that, for most of them, outsourcing our examinations to another person would not prevent their interference. It will be true of any attempting to examine another person that:

- Her occurrent thoughts and feelings will often cloud her examination.
- Her examination will sometimes get derailed because she will end up identifying the issues which are are easiest to put into words even if these are not the most significant for what she needs to report.
- She will often respond by appealing to idiosyncratic or shared cultural theories which describe how or why people in general (or this person in particular) typically feel, think, or act, theories that might not explain for how or why they actually felt, thought or acted.

63. It might be worth saying that a person cannot examine herself objectively through and through. If the project is one of self-examination, then the person can never be kept entirely out of it. There has to be a moment when she has to be acquainted with the results of the examination and in this moment it is always possible that her unconscious processes interfere and distort how this information is received and assimilated.

64. Wilson suggests this in the book's first chapter (Timothy D. Wilson 2002, pp. 1–16) and explicitly defends it in Timothy D. Wilson and Dunn 2004, pp. 494–504.

- She will rely on faulty observations of correlation or covariance between antecedent and subsequent events.[65]

There does not seem to be any principled reason to think that these unconscious interferences will be reduced when they are outsourced to a third person. Anyone is likely to end up relying on what is currently on her mind, on what it is easiest to put into words, on a theory about how people behave, or on faulty observations of covariance. Outsourcing your investigations to a third party does not seem efficacious to prevent these types of intrusions. It is useful to ask a third person to help you examine yourself when you are blinded by your hubris, when you are motivated to have an overly positive view of yourself, or when you are repressing your mental states because they bring about anxiety or psychic pain. Outsourcing self-examination works best when the issues that interfere with the acquisition of self-knowledge, issues concerned your self-worth or your self-image. But these are the types of cases which Wilson is not interested to discuss.

Ethical issues have to do with the question "how should I live?" This question, in turn, is internally related with the question "Who should I be?," a question that is deeply entangled with issues about our self-worth and our self-image. Arguably, then, examining mental states that have to do with our ethical development will be difficult to examine. As a consequence, third-personal methods of self-examination might often be more reliable than first-personal methods. If one believes that first-personal self-examination is not reliable way to know ourselves in cases where issues of self-worth or self-image are not at stake, one needs to provide an explanation for why first-personal self-examination is less reliable than third-personal self-examination. Wilson does not provide us with a satisfactory one.

65. And it might be relevant to mention that these faulty observations of covariance are caused, according to Wilson, not by hidden motives, but to the fact that our conscious system is "notoriously bad at detecting correlations between two variables. In order to detect such relationships, the correlation has to be very strong, and people must not have a prior theory that misleads them about this correlation."Timothy D. Wilson 2002, p. 62

In the book, Wilson attempts to offer a criticism of first-personal methods of self-examination. But what he actually ends up doing is showing that, when we try to know ourselves, both first- and third-personally, different types of unconscious mechanisms get in the way of our own investigations. Given the type of mechanisms that he discusses, however, it is not clear that he has the resources to show that the latter method of self-examination is more reliable than the former.

III.4.3 First-Personal Grounds and One's Relationship With Them

I would like to conclude the chapter by bringing up a further reason which explains why it is natural to think that third-personal self-examination is more reliable than first-personal self-examination, an idea which arises from an important asymmetry between the way in which another person confronts one's first- and third-personal self-ascriptions..

When we examine ourselves third-personally the grounds that justify our conclusions are grounds to which any other person would be entitled to appeal if she was to justify them. Because anyone is entitled to appeal to them it is also possible for anyone to examine these grounds and to assess the extent to which they provide a sound justification for the conclusion. By contrast, the grounds which justify my first-personal claims might not be available to anyone else other than myself. In fact I might not have any grounds like the grounds that a third person would require to have. When I know a mental state first-personally, I am entitled to self-ascribe this mental state even if I am unable to cite evidence in support of it.[66] This establishes, as it were, an inherent dogmatism in first-personal self-

66. Different epistemic theories of self-knowledge will account differently for the grounds that justify my self-knowledge. Inferentialists will tend to say that my first-personal self-examination is, ultimately, grounded on inferences from my behavior. According to inner-sense theorists, it will be grounded on a quasi-perceptual perception of my mental states. Idealists like Sebastian Rödl will say that they are grounded on spontaneous knowledge, i.e. knowledge one has of an object just by being such object (Rödl 2007). Some expressivists as well as proponents of the agency model of self-knowledge are likely to hold that my first-personal claims are not epistemically grounded.
My claim, here, is not that self-knowledge acquired through first-personal self-examination is groundless, but that the grounds that are typically associated with ways of knowing that rely on perception, testimony or

examination. First-personal self-examination entitles me to claim that I have a mental state without requiring any third-personal grounds justifying my claim. This is not something that is possible when a third person is making a claim about myself. I am entitled to ask for her grounds. And her failure to provide them weakens her claims in ways that my lack of grounds does not weaken my claims when they are the product of first-personal self-examination.

When I conclude, through first-personal self-examination, that I hold a mental state, I am entitled to not provide grounds for my self-ascription. As a consequence others will not able to challenge the grounds on which my assertion is based. If I say "I am angry," it will be possible for anyone else to contradict my self-ascription. But to do this they will need to appeal to grounds such as my behavior. But it is worth keeping in mind that even upon learning about these grounds, and even if these grounds are compelling, it is still open for me to refuse to accept them as justifying the denial of my claim. I am entitled to say: "It is true that I have been behaving as though I was not angry... but the fact of the matter is that I am angry."

It is not uncommon to find ourselves challenging someone's assertion by appealing to different grounds from those that justify this person's assertion. You might tell me that there are turtle eggs in the river because you saw them; I might challenge you by replying that there are no turtle eggs in the river because the natives have never seen turtles in this river. Here we are making contradictory assertions by appealing to different grounds. Your assertion is grounded on your perception, mine on the authority of the natives. If it is important to settle this disagreement, you can ask me to corroborate your grounds, say, by asking me to come with you to the place where you saw the eggs and see them for myself. I, in turn, can share with you the grounds that have convinced me that the authority of

---

inference might not be available to the person examining herself first-personally. This is, in fact, one of the central features of first-personal self-examination, that it proceeds without an explicit attempt to ground the person's knowledge in the way that a third person typically would. It is a form of self-examination that is subjectively felt to be immediate and private, a form of examination that we intuitively think is exclusive to the subject, not available to anyone but ourselves. Any theory of self-knowledge should have room to account for this phenomenon.

the natives is trustworthy, grounds which you can, at least in principle, also corroborate by yourself.

This second part of this dialectic is not available when our disagreement is about a self-ascription that I attribute to myself on first-personal grounds and you attribute to me on third-personal grounds. You can, certainly, contradict my self-ascription. You can even get me to see that your grounds are strong and that mine are weak (say, by showing me how conclusive your evidence is and by appealing to my poor track record of reporting on this issue). But if I am entitled to make claims about myself first-personally without any grounds, then you cannot contradict my grounds.

The evidence at stake in self-knowledge acquired through third-personal self-examination is evidence that I am supposed to share with others and, therefore, evidence that others can help me assess. The evidence at stake in self-knowledge acquired through first-personal self-examination is not of this sort. This asymmetry brings out that there is a kernel of truth in the idea that first-personal claims are irrefutable. First-personal claims are irrefutable to the extent that it is not possible, for anyone else, to examine the first-personal grounds when I refuse to offer them. This irrefutability makes first-personal claims, at least from a certain perspective, dogmatic. This dogmatism, in turn, explains why we are inclined to take our first-personal claims for granted, why we might be less willing to scrutinize and revise them. And this, in turn, might lend more credence to the thesis that third-personal self-examination can be more reliable that first-personal self-examination.

III.5 What We Learn from the Experiments

In this chapter I established some conceptual distinctions that allowed us to clarify and assess some of the main theses defended by Wilson in his work. I argued, *contra* Wilson, that self-examination in general and first-personal self-examination in particular are not as unreliable

as he takes them to be. His theoretical arguments justifying this view are inadequate. They trade on equivocating between higher-order processes and sub-personal processes and fails to register that a process that is unconscious here and now might become conscious in the future. The empirical evidence does not support it. The ecological validity of the studies is questionable and they fail to establish the causality at the right level of description. Finally most of the mechanisms that, according to Wilson, make first-personal self-examination unreliable are mechanisms that are also liable to corrupt third-personal methods of self-examination. What is more, Wilson himself recognizes that there are occasions where it is precisely the cultivation and strengthening of our capacity for first-personal self-examination that improves our ability to know ourselves.

Despite all of this I concluded the chapter by providing a different argument to support what I take to be a kernel of truth in Wilson's view, namely, that in the case of ethical development, particularly in the case where issues of self-worth or self-image are at stake, there are occasions where third-personal self-examination will tend to be more reliable than first-personal self-examination. In developing this claim, however I challenged Wilson, arguing that what made third-personal self-examination more reliable in these cases had to do with the kinds of motivated irrationality discussed in traditional accounts of self-deception and not with the mechanisms associated with the "adaptive unconscious."

The empirical phenomena to which Wilson appeals are disturbing in a number of ways. They point to our fallibility to make adequate judgments caused by the influence of minor situational cues that lead us astray. And although at some level of description it might be true that the experiments bring out problems with our capacity to know ourselves, in particular with our incapacity to properly track important influences on our behavior, they fail to establish that we are not very good at knowing our higher mental processes.

But perhaps the most important thing that many of these experiments reveal is that subjects come up with confabulated responses with some regularity and that they are ut-

terly blind to this fact,[67] certain of the correctness of their reports, and firmly committed to them. When we rationalize or confabulate we are usually completely unaware that we are doing it.[68] What some of this date suggests, as Peter Carruthers puts it: "is that our common-sense belief in the existence of introspective access to judgments and decisions is without epistemic warrant, and that sometimes, at least, our access to our own attitudes is actually (but unconsciously) interpretative."[69] One of the most valuable recommendations that Wilson makes in his book is that we should be less confident about the reliability of our self-examination and, consequently, more humble about our capacity to provide accurate reports about ourselves.[70] His discussion should serve as a valuable reminder that triangulating information about oneself that comes from different sources is an excellent way to improve the accuracy of our self-investigations.

The conceptual work done so far and the way in which this work has deepened our understanding of the key terms involved in this discussion, equip us to move firmly into the next chapter. I will argue that, regardless of the fallibility of our capacity for first-personal self-examination, coming to know our mental states first-personally is ethically significant. Wilson has difficulty finding conceptual space to accommodate this view because he works with an utterly contemplative conception of self-knowledge. I will show that, even if we concede that there are cases where our capacity for first-personal self-examination is more fallible than objective methods for self-examination, this is not a sufficient reason to think that coming to know our mental states third-personally is ethically superior. In fact, I will show that Wilson's ideal self-knower is alienated from herself and that this makes her ethically handicapped, interfering with her ability to live an ethically excellent life.

67. Timothy D. Wilson 2002, p. 167.

68. Timothy D. Wilson 2002, p. 168.

69. Carruthers 2010, p. 87.

70. Timothy D. Wilson 2002, pp. 112-3, 168.

IV

Endorsed First-Personal Self-Knowledge

There are myriad unconscious mental states that shape our behavior. Wouldn't it be nice if there were an app, an "Inner Self Detector," that could just provide us with reliable and accurate reports about these opaque mental states? Most of us would think so. Such an app would help us identify aspects of our life that interfere with our ethical development and it would thereby help us to make better ethical choices.[1]

But what do we mean by acquiring self-knowledge? Do we just mean acquiring reliable and accurate information through an app like the Inner Self Detector? In this chapter and the next I will argue that the answer to this question is "no." The aspiration to achieve self-knowledge, if it is guided by an aspiration to become virtuous, is not merely an aspiration to come to know certain facts about ourselves; it is also an aspiration to engage with what is known in a manner that allows the ethical pilgrim to shape it. This in turn ultimately enables

1. I am borrowing the name "Inner Self Detector" from Timothy Wilson's book, *Strangers to Ourselves* (Timothy D. Wilson 2002, pp. 3, 120). In the book Wilson suggests that psychologists are slowly developing such a tool: "True, we do not yet have an Inner Self Detector, but increasingly sophisticated techniques are being developed, such as measures of the neurological correlates of emotion and affect" (Timothy D. Wilson 2002, p. 120).

the ethical pilgrim to unify mental states that conflict with one another into a coherent whole and to be a properly self-determining rational creature.

In the last two chapters of the dissertation I intend to bring out the ethical significance of the distinction between what I will call first- and third-personal self-knowledge. Although a version of this distinction has been at the heart of a number of important debates within epistemology, its ethical significance has seldom been explored or acknowledged.[2] In showing that this distinction is ethically significant the dissertation puts the work that epistemologists have done on self-knowledge and first-person authority in contact with the work that ethicists have done on moral education and ethical development.

Within epistemology the main question about self-knowledge is the question about how it is that one has what seems to be immediate, private and authoritative knowledge of one's mental states. My project does not engage with the specific issue that epistemologists discuss under the rubric of self-knowledge. I am not interested in offering a metaphysical account of the mind that explains how we can know our mental states in what seems to be a privileged, immediate or authoritative way. Rather, my dissertation approaches the topic of self-knowledge from a practical perspective. It aims to understand the ethical significance of different ways in which a subject can relate to her knowledge of her mental states. For my purposes in the dissertation it will not be important to determine how the subject ultimately comes to know about them. One of the virtues of this approach is that it is compatible with most theories of self-knowledge defended by epistemologists.

The first three sections of the chapter are concerned with clarificatory work that will provide the main conceptual framework required required to develop the arguments in the rest of the dissertation. In the first section I define and clarify the notions of first- and third-personal self-knowledge. In the second section I articulate the relationship between

2. Some epistemologists have actually asserted that these distinctions are not ethically significant (see, for instance, Carruthers 2011, p. xi). Among the epistemologists who have tried to engage with the ethical dimension of self-knowledge are: Richard Moran 2001; McGeer 2007; Cassam 2014.

first/third-personal self-*examination* (discussed in the previous chapter) and first/third-personal self-*knowledge*, (discussed in the first section of this chapter). Developing my argument requires that I distinguish between more than one kind of first-personal self-knowledge.

In examining the place that first-personal self-knowledge has in ethical development I have come to see that a proper understanding of it requires that one distinguishes between at least two different kinds of first-personal self-knowledge. In the third section of this chapter I define and clarify two types of first-personal self-knowledge that will be important for the rest of the argument: "endorsed self-knowledge" and "merely-expressive self-knowledge."

The rest of the chapter is focused on examining the role that endorsed self-knowledge plays in ethical development. The central intuition guiding my argument is that we are not merely witnesses of our mental states; we are also capable of knowing our mental states in such a way that our judgment about the merits of holding them can alter them.

In section 4 I argue, *contra* Wilson, that endorsed self-knowledge is ineliminable and indispensable in our everyday interactions. In doing so, I show that it is unfeasible for Wilson to think that one could live according to the instructions provided by the Inner Self-Detector. In sections 5 and 6 I argue that endorsed self-knowledge plays a central role in the virtuous life. Lacking endorsed self-knowledge undermines your capacity to make up your mind in a properly rational way, to live your life within a long-term perspective that understands it as the answer to the overarching question "how should I live?" and to allow you to reflect on this answer in the company of others. Possessing endorsed self-knowledge of your mental states, on the other hand, reflects your unity as speaker, reasoner and doer. It follows from these arguments that Timothy Wilson's ideal self-knower—the person who leads her life guided by third-personal self-knowledge—is alienated from herself in a way that is detrimental to her ethical development and her well-being.

IV.1 Defining First- and Third-Personal Self-Knowledge

The mental states of a person manifest externally in myriad ways. Take the case of anger. When you are angry, your face might turn red, your respiration probably becomes faster, and you may raise your voice or slam a door. All of these behaviors are expressions of anger. Some of these can be found in non-rational animals. Human beings, unlike brutes typically have an additional way to express their anger: they can *express* their anger through a self-ascription. Saying (or yelling) "I am angry!" is a self-ascription that often serves both to report your anger and to express it.

Your self-ascriptions are not always expressions of your anger. There are cases where you are at a certain distance from your anger and you can merely report on it as though you were a mere witness of it. In this kind of case, your self-ascription is merely a report of your anger, a report that does not express such anger. Think of the case where you come to know that you are angry at your partner through the evidence provided by your therapist. You might not take yourself to be angry at him, you might not acknowledge such anger. But your therapist is right and the behavioral evidence is difficult to dispute. If you come to believe what your therapist tells you, namely, that you are angry at your partner, you come to know something about yourself, you come to know that you are angry at your partner. You are, however, alienated from your knowledge. Although you can self-ascribe your anger, the self-ascription "I am angry" is merely a report of your anger, it is not an expression of it.

This prepares the ground for my definition of first- and third-personal self-knowledge.

First-personal self-knowledge: I will say that you have "first-personal self-knowledge" of a mental state M when your capacity to self-ascribe M is such that the mere ascription

of M is, at the same time, a (knowledgeable) report of M *and* an expression of M.[3]

Third-personal self-knowledge: I will say that you have *"third-personal self-knowledge"* of a mental state M if your capacity to self-ascribe M allows you only to make a (knowledgeable) report of M but not to express M in the mere self-ascription of M.

(There will be occasions in the discussion where I will want to leave open whether the self-knowledge that I am discussing is first- or third-personal. When this is the case I will simply talk about "self-knowledge" without qualification).

It might be tempting to think of the distinction between knowing that I am angry first- or third-personally as the distinction between yelling "I am angry!!!" and neutrally reporting "I am angry." This might be a good way to get a first, intuitive grasp, of the distinction between these two types of self-knowledge. But it is not an accurate one. This intuitive portrayal could perhaps best be described as a caricature of the phenomenon. It is a picture that gets at some of the central features of the phenomenon but which is lopsided, exaggerated and imprecise. First, it is possible to *express* one's anger in a neutral tone of voice without, so to speak, "exclamation marks." Perhaps the social circumstance is such that the "exclamation marks" are inappropriate and you opt to express your anger without them. Or perhaps the anger is a long standing one that you've held for many years, and you are at a point where you just express it matter-of-factly.[4] On the other hand, you might be yelling "I am angry!!!" in an emotively charged way, even when your self-ascription is third-personal. This can be so, for instance, when your yelling is caused by a different emotion that happens to be expressed in this self-ascription.[5]

3. c.f. Finkelstein 2003, p. 120.

4. The fact that you can express your anger matter-of-factly illustrates that an emotion like anger need not always be something that you "feel," a consideration which, in turn, gives support to the view that emotions like anger are cognitive mental states, closer in nature to beliefs than to sensations.

5. It can even be the case that this self-ascription might be an expression of this anger, but an expression of which the person is unconscious. I will return to this last case shortly.

Throughout the dissertation I will be using the notion of "expression" to characterize acts that are *true* and *genuine* manifestations of the person's mental states. It is possible to feign anger, and in ordinary language we sometimes refer to this kind of act as an act where the person is "expressing anger." My usage of "expression," however, does not allow for this. As I will use the term, feigning anger is not an expression of anger (it is, rather, a "feigned expression" of anger).[6]

In common parlance one might say that when a person self-ascribes a mental state, regardless of how she self-ascribes it, she is expressing this mental state. I am not using expression in this wide way. I use it, instead, to characterize behaviors which manifest these mental states directly. Thus, to express a mental state is not merely to speak about it or to utter a report about it. According to my usage, when you report, after talking to your analyst, that you are afraid of a certain person, your report might *express that* you are afraid, but it does not *express your fear*. For a mental state to be expressed in a certain behavior, it has to be a manifestation of such mental state, it has to be directly present in it.

My definition of first- and third-personal self-knowledge is highly indebted to David Finkelstein's definition of conscious/unconscious states of mind.[7] My proposal, as I mentioned in the opening of the chapter, is meant to be neutral with respect to the theories of self-knowledge proposed by epistemologists. My account is compatible with many other theories of first-person authority, provided that 1) these theories are able to make sense of the distinction between first- and third-personal self-knowledge, and 2) these theories can account for the non-accidental connection between expressing a mental state in a self-ascription

6. It has been well documented that feigning the expression of an emotion sometimes leads to the actual development of such emotion. When this happens the person is no longer feigning this emotion. Some readers might think that this kind of case suggests that it is not possible to neatly distinguish between feigned expressions and genuine expressions. I don't think that this is right. The fact that a feigned expression of an emotion might help to develop such emotion should not lead us to relinquish the distinction between feigned expression and genuine expression. Even if there is a causal correlation between feigning a mental state and developing this mental state, this correlation need not entail that there is a conceptual correlation between feigning the expression of a mental state and genuinely expressing it.

7. See, in particular, Finkelstein 2003, p. 120 and Finkelstein 1999.

and knowing that you hold this mental state.

IV.1.1 Clarifying the Definitions

In what follows I will clarify some of the individual elements which compose my definition of first- and third-personal self-knowledge. These clarifications will sharpen these definitions and, in doing so, will solidify our understanding of them. Readers who are impatient to get to the discussion about ethics might want to skip ahead to section IV.3.[8]

Precluding a Non-Accidental Connection

I defined first-personal self-knowledge as the capacity to self-ascribe a mental state such that the *mere* self-ascription of S does not just report S but also expresses S. The qualification that I just emphasized (i.e. saying that the *mere* ascription expresses your mental state) is meant to exclude cases where it might turn out that, by accident, the mental state that is expressed happens to coincide with a mental state that is known only third-personally. Take the case of a person who knows, third-personally, that she is angry but who self-ascribes her anger in a tone of voice that, unbeknownst to her, expresses her anger. In this case, even though her self-ascription expresses her anger, her knowledge of such anger is still third-personal. And this is so because the anger latches on to the person's self-ascription without her self-conscious recognition of this fact. In a case like this the anger's self-ascription is an expression of it, but the connection between the anger and the expression is accidental; the anger is not expressed *in* the self-ascription but *through* it. I stipulate that the mental state is expressed *merely* through a self-ascription to ward off these kinds of accidental cases. In the particular case that we are discussing, what expresses this person's anger is not the *mere*

8. Those who skip to section IV.3, however, should note that in section IV.2 I establish that, when successful, first-personal self-*examination* leads to first-personal self-*knowledge* and third-personal self-*examination* will lead to third-personal self-*knowledge*. Although this should not be a surprising conclusion, I do elucidate what warrants it.

self-ascription but, rather, the tone of voice that latches onto such self-ascription.[9]

Why Call Them "First-Personal" and "Third-Personal"

The labels "*first*-personal" and "*third*-personal" are meant to make perspicuous the fact that the former kind of self-knowledge is available *only* to the subject of the mental state while the latter is available, at least potentially, to a third person. What secures the first-personal dimension of first-personal self-knowledge is the fact that only the subject can *express* her mental states by self-ascribing them. This is so because it is only the subject of a mental state who can express it. When you are speaking about the mental state of another person, you can report on her mental state, but you cannot express it. My cries, sighs, and frowns express my pains, frustrations and discomforts. And even though you can empathize with me, and cry because of my pain, when you do so you are not expressing my pain or frustration but yours.

First/Third-Personal Self-Knowledge as Capacities

I have defined first- and third-personal self-knowledge as a *capacity*. This entails that having first- or third-personal self-knowledge of M does not require one to effectively self-ascribe M. All one needs is to have the capacity to do so. It is worth pointing out, however, that this is a capacity that should be available to the person here and now, a capacity that she should be able to deploy whenever she wants.[10]

9. This is a move that I am borrowing from Finkelstein 2003, p. 120, 1999, p. 94.
Where I refer to this definition I will usually omit explicit references to this qualification for simplicity.

10. In other words, it is what in Aristotelian jargon one would call a capacity in first-actuality.

Two Ways of Being Conscious

In a previous chapter I defined a *conscious* mental state as a state that a person could (genuinely) self-ascribe (II.2). It follows that first- and third-personal self-knowledge are a species of conscious states of mind. It is worth noting, however, that it might be natural to characterize the mental states of a person that she knows first-personally as "conscious" and those which she cannot know first-personally as "unconscious." The apparent tension between these two intuitions can be resolved by establishing a distinction between two ways of being conscious of our mental states. We often say that the person who comes to know that she is angry through the evidence provided by the analyst is conscious *of* her anger. And by saying this we tend to mean that the person is aware of her anger, that she knows about it. But we tend to distinguish these cases from the case where the person is conscious*ly* angry. In this latter case, the person does not merely know that she is angry. She actually stands in an intimate relationship to her anger, a relationship that allows her to have the capacity, not merely to know or be aware of the fact that she is angry, but to express her anger in a self-ascription.[11]

A "Knowledgeable" Report

I defined first- and third-personal self-knowledge of a mental state M as a capacity to make a "knowledgeable" report of M. This was meant to secure that the report is actually a case of self-knowledge. Cases where I make a guess and happen to hit on the truth by accident are not cases of knowledge. It is to avoid characterizing these kinds of cases as cases of self-knowledge that I defined first- and third-personal self-knowledge as the capacity to make a *knowledgeable* report of M.

11. For a fuller explanation of this distinction see Finkelstein 1999, pp. 80–1, 2003, pp. 114–6.

The Distinction Beteween First/Third-Personal Self-Knowledge Is Not About the Content

The difference between knowing a mental state first- or third-personally has nothing to do with the content of what one knows. Suppose that I know that I am angry at my partner. Knowing this first-personally and knowing it third-personally entails knowing the same thing, namely, that I am angry at my partner. The difference between knowing it first- or third-personally has to do, not with *what* I know, but with *how* I relate to what I know. Only when I have first-personal self-knowledge am I able to use my report of M also as an expression of M. When I only have third-personal self-knowledge it is not possible for me to do this.

Second-Personal Self-Knowledge

My definition of first- and third-personal self-knowledge is exhaustive. Whenever you know your mental state you know it either first- or third-personally. I have often been asked about the place for second-personal self-knowledge in my account. My answer to this question depends on what the person asking means by "second-personal self-knowledge." Second-personal self-knowledge would be a species of first-personal self-knowledge if the second person somehow shares the subjectivity of the person. Only thus could she have the capacity to express the first person's mental state in a self-ascription (a self-ascription which would, thereby, also be a self-ascription of the mental state of the first person). However if the second person, however sympathetic or compassionate, cannot express the first-person's mental state in a self-ascription, then the second-personal knowledge that she could provide would amount to a species of third-personal self-knowledge.[12]

12. I find it plausible to think that, within certain forms of psychotherapy, the therapist might have first-personal self-knowledge of the mental states of the client. Although I find this line of investigation attractive, it is very contentious and brings with it a number of complications which lie beyond the limits of this project. Thus I will be working with cases where the subject of the mental state corresponds to the individual person

A Caveat

It is worth explicitly mentioning that if the importance of coming to know one's mental states was only a matter of acquiring information about these mental states, regardless of how one relates with this information, then there would not be any difference between third-personal self-knowledge and first-personal self-knowledge. My whole thesis, then, depends on denying this antecedent, an antecedent that Wilson's thesis seems to simply take for granted.

IV.2 First-/Third-Personal Self-Examination; First-/Third-Personal Self-Knowledge

How do the two methods to investigate ourselves that I discussed in the previous chapter, first- and third-personal self-*examination*, relate to the two forms of knowing that I am describing in this chapter, i.e. first- and third-personal self-*knowledge*?

In this section I will argue that there is an internal connection between first-personal self-examination and first-personal self-knowledge on the one hand, and between third-personal self-examination and third-personal self-knowledge on the other.

IV.2.1 The Internal Connection Between First-/Third-Personal Self-*Examination* And First-/Third-Personal Self-*Knowledge*

Let me start by justifying that first-personal self-examination typically leads to first-personal self-knowledge. If you think of any case when first-personal self-examination leads you to make a report on one of your mental states, you will find that this report will, typically, be

who self-ascribes it. I'd like to thank Martha Nussbaum, Francey Russel and Nancy Sherman for pressing me on this issue.

also an expression of such mental state. One might think that this is even one of the main hallmarks of first-personal self-examination.

I mentioned in the previous chapter (III.4.3) that the self-ascriptions that you make as a result of first-personal self-examination are not grounded in any of the grounds to which a third-person would need to appeal if she wanted to ascribe a mental state to you in an epistemically responsible way. But these self-ascriptions do not appear arbitrary. Despite their lack of third-personal grounds, the subject reporting her mental states in this way takes herself to be entitled to make these self-ascriptions. I mentioned that first-personal self-examination is a form of self-examination that *is taken to be* direct, private and privileged. It is a form of examination that we intuitively think is exclusive to the subject, available to no one other than the subject herself (III.2.2). My suggestion is that the entitlement that you take to have to self-ascribe a mental state after first-personal self-examination depends on the fact that you are expressing your mental state in a self-ascription. If you are asked, "But how do you know that you hold mental state M?" you do not cave in and apologize for not having grounds to self-ascribe M. Instead you respond as though the person is joking or not really understanding you. If the person insists, requesting you to provide the kinds of grounds that a third person would be required to provide, you will complain and argue that your interlocutor is failing to see that you have a certain authority to speak about your mental life, that you are in a privileged position to know those mental states, an authority that comes from the peculiar relationship that you have to them. This alleged authority that you take yourself to have comes from the fact that your self-ascription is also an *expression* of this mental state. From the fact that you can express a mental state it follows that you are entitled to self-ascribe it. When first-personal self-examination leads you to make a self-ascription, you are not guessing or coming up with this ascription out of the blue. First-personal self-examination leads you to express a mental state in a self-ascription. Because you are able to express a mental state in a self-ascription, you are entitled to make

knowledgeable reports on it. Thus, *first*-personal self-examination, when successful, leads to *first*-personal self-knowledge.[13]

When third-personal self-examination leads to self-knowledge of a mental state M, you come to know that you hold M in the way that a third person would come to know it, namely, through inference, testimony or perception. Because these third-personal grounds are explicitly available to the person, she will be able to make claims about herself based on them. If you are merely reporting on a mental state, but your report is not an expression of your mental state, then your report, to be knowledgeable, will need to be grounded on third-personal grounds.

The internal connection between first-personal self-examination and first-personal self-knowledge on the one hand, and third-personal self-examination and third-personal self-knowledge on the other, entails that we can think of first- and third-personal self-knowledge in the following way. Your self-knowledge of a mental state is "third-personal" when you explicitly take it to be grounded in a way of knowing that is available, at least potentially, to another person (or, to be more precise, to another subject). Your self-knowledge of a mental state is "first-personal" when you take it to be explicitly grounded in a way of knowing that is not available to anyone other other than yourself (or, to be more precise, to anyone other than the subject of the mental state).

13. One might think that, in saying this, I am inadvertently committing myself to certain expressivist theories of first-person authority. What I am committing myself to here is to the view that any theory of self-knowledge has to make room for the fact that the capacity to express your mental state in a self-ascription entitles you to make knowledgeable claims on your mental states. Other epistemic theories of self-knowledge, such as inner-sense theories or inferentialist theories, might be able to show that this capacity relies on a more primitive capacity to know our mental states, a capacity reliant on your inner sense or on (perhaps tacit) inferences and interpretations from your observed behavior.
Some schools of thought within epistemology, particularly those influenced by empirical research, might characterize the internal relationship between first-personal self-examination and first-personal self-knowledge as the result of a certain reliability built into our mental architecture. Other schools will characterize this relationship as dependent on a constitutive conceptual relationship. According to both schools of thought, however, first-personal self-examination, at least when successful, will typically lead to the capacity to express your mental state in a self-ascription.

IV.2.2 *Third-/First-* Personal Self-Knowledge and *First-/Third-* Personal Self-Knowledge

Even if first- and third-personal self-examination, when successful, lead to first- and third-personal self-knowledge, respectively, it is worth pointing out that there are cases where third-personal self-examination can lead to first-personal self-knowledge.

This is a conceptual possibility that Timothy Wilson misses and which leads him to mistaken conclusions about the scope and nature of self-examination and self-knowledge. Even if the typical outcome of a successful third-personal *self-examination* is the achievement of third-personal *self-knowledge* this need not mean, either that this is what always will happen, or that this is what is meant to happen. To see this it should suffice to realize that there are plenty of cases where our third-personal self-examination helps us to reconnect with our feelings and express them adequately. Think again of the example I discussed above. It can happen (and in fact it often happens) that all that it takes for me to acknowledge my anger is to be told about it. My therapist might make me realize that I am angry at my partner. But this realization, initially based on third-personal self-knowledge, might be all that it takes for me to identify my anger 'from the inside,' to be able to express it.

By the end of the dissertation I will have argued, not merely that *third*-personal self-examination might happen to lead, in some cases, to *first*-personal self-knowledge but that, within many domains, in particular when one is thinking about ethical development, the ultimate aim of self-examination, even of *third*-personal self-examination is the acquisition of *first*-personal self-knowledge. Showing this will help us understand why transitioning from this third-personal way of examination to a first-personal way of knowing constitutes one of the cornerstones of some therapeutic approaches to mental health. In particular it will allow us to understand what is at the root of Freud's idea that third-personal self-examination and

third-personal self-knowledge should be cultivated to attain first-personal self-knowledge.[14]

But while third-personal self-examination might lead to first-personal self-knowledge, it is not at all clear that the converse is true. When you have first-personal self-examination you are not attempting to seek any of the grounds that would entitle a third person to attribute a mental state to you in an epistemically responsible way. Thus, it would be quite an accident if the person, who is not seeking third-personal grounds, ends up acquiring them.

This brings out an asymmetry between first/third-personal forms of examination and first/third-personal forms of knowledge that is worth explicitly highlighting. Typically, first-personal self-examination leads to first-personal self-knowledge and third-personal self-examination leads to third-personal self-knowledge. However, first-personal self-examination does not lead to third-personal self-knowledge, third-personal self-knowledge, not only may lead to first-personal self-knowledge but in some cases ought to lead to third-personal self-knowledge.[15]

IV.2.3 Are First-/Third-Personal Self-Knowledge Exclusive Categories?

Just like first- and third-personal ways of examination can work in tandem, it is also possible to acquire self-knowledge that is grounded on a combination of first- and third-personal considerations. The example that I discussed in the previous chapter can be redeployed to illustrate this. In that example a friend asked Hellen whether she loved her new boyfriend, Harry. Hellen replied: "Well, I feel comfortable when I'm with him, and I am really attracted to him. But he talks too much about his therapy sessions and I don't always like his politics. And you know how I can't stand being around people when they are sick? Well... When

14. For evidence that Freud thought that third-personal self-examination was meant to lead to first-personal self-knowledge see Freud 1995a, pp. 141-2, Freud 1980, p. 155 or Freud 1963, pp. 281, 437.

15. I provide a specific example of this towards the end of the chapter (IV.8.1)

he had the flu I stopped every evening to bring him food and check up on him. So, yes, I suppose I love him."[16] In the case that I am imagining here, Hellen is at a moment in her relationship where she is gradually coming to acknowledge that she is in love with Harry. The feelings for him, however, are not yet fully fledged so she needs to complement her capacity to express them with third-personal evidence. In a case like this, her self-knowledge relies on a combination of first and third-personal grounds, that is to say, on a combination of her inchoate capacity to express her love for Harry and behavioral evidence that provides some grounds to think that she loves him. As I am conceiving the example, neither of these two types of grounds, on their own, are sufficient to entitle her to know that she loves Harry. Her capacity to express her love is inchoate and her behavioral evidence is far from conclusive. It is the combination of both that warrants considering her self-ascription as a case of self-knowledge.

There are also cases where the person can have, at the same time, first-personal self-knowledge and third-personal self-knowledge. To see this, it suffices to realize that there are many cases where we have the capacity to express a mental state in a self-ascription but where we also have the capacity to report on this mental state as an informed witness of it, providing third-personal grounds that justify self-ascribing it.

IV.2.4 An Idealized Account?

When I have discussed third-personal self-knowledge I have often portrayed the knowledge you have as a piece of information about someone who just happens to be you. Nancy Sherman has suggested that even when you come to know facts about yourself in this third-personal way you are not relating with this knowledge merely in the way that a third person does. Even an externalized perspective on myself, she suggests, will affect how I conceive

16. I am borrowing this example from Finkelstein 2003, p. 123.

of myself and how I behave.[17] I agree on this. Knowing something about myself, even if it is known third-personally, is likely to have an impact on myself, on how I relate to this knowledge and on my capacity to express it. I am even willing to grant that it is perhaps never possible to know one's mental states in an entirely third-personal way.[18] I find it plausible to think that, when I know that I have a mental state, even if this knowledge is third-personal, I might have at least a very minimal and inchoate capacity to express it in a self-ascription. If this is so, then the way in which I characterize the distinction between first- and third-personal self-knowledge on which I am relying is somewhat idealized because there would not be a pure case of third-personal self-knowledge.[19] I don't think that this fact poses any serious threat to my account. Even if it were true that we can never know ourselves in a purely third-personal way, even if self-knowledge always entails a capacity, regardless of how minimal and inchoate, to express our mental states in a self-ascription, the distinction between these two forms of self-knowledge is still of fundamental importance to conceptualize cases where one's capacity to express one's mental state is more developed than in other cases where such capacity is quite minimal. In fact, it is this distinction which allows us to cash out the idea that third-personal self-knowledge provides an "externalized perspective" to begin with.

In the last two sections I defined first- and third-personal self-knowledge, explained the different elements of these definitions in detail and articulated how these forms of self-knowledge relate with the two forms of self-examination discussed in the previous chapter. In the next section I will need to establish one more distinction that is central to the dissertation, the distinction between endorsed first-personal self-knowledge and merely-expressive first-

17. Nancy Sherman. Personal Communication, April 01, 2015.

18. Except, of course, in the case where the person's true self-ascription is not *de se*. That is to say, the self-ascription is true despite the fact that this person does not recognize that it being predicated about herself (Castañeda 1966).

19. I am bracketing, of course, cases of self-knowledge where the self-knowledge is not *de se*. It is quite likely that these cases will allow for pure cases of third-personal self-knowledge.

personal self-knowledge. In the following section I will define these two kinds of first-personal self-knowledge and develop a terminology to speak about them. This will allow me to argue, in the rest of the chapter, that the ethical pilgrim cannot merely aspire to know herself third-personally.

IV.3 What Is Endorsed First-Personal Self-Knowledge?

IV.3.1 Defining "Endorsed Self-Knowledge"

In most cases, the person's utterance "I am angry!" will express, not merely the person's anger, but also her endorsement of such anger, it will express the fact that she takes her anger to be warranted. An emotion like anger, however, is not always pliable to reason and there will be occasions when a person will be angry despite her own recognition of the emotion's irrationality. In these cases in which an emotion is recalcitrant to reason, the person can still express her anger in a self-ascription, even though such expression will not be also an expression of her endorsement of her anger.

We can think of these two ways to relate to your anger as two ways to relate with your self-knowledge, two types of first-personal self-knowledge. These two ways to know your mental states are both ethically significant (although in different ways and for different reasons) and suggest the following definition:

Endorsed self-knowledge: To express your endorsement of a mental state, as I mentioned above, requires the person to take the mental state to be warranted. I will refer to this form of first-personal self-knowledge as "*endorsed first-personal self-knowledge*" (or "*endorsed self-knowledge*" for short). I will also say, following the convention I used earlier (sec IV.1.1), that the person is "consciously endorsing" her mental state.

Merely-expressive self-knowledge: To distinguish this case from the case where the person has first-personal self-knowledge but lacks endorsed self-knowledge I will say that the person has "*merely-expressive first-personal self-knowledge*" (or "*merely-expressive self-knowledge*" for short).

To say that the person "expresses her endorsement" might be read in two ways. And it is important to disambiguate to properly understand how I will be using this locution. On the one hand, one might think that a person endorses a mental state when she can clearly identify and convey the warrants that justify it. Take the father who has received a telegram informing him that his daughter was killed in war. He can say "I believe that my daughter was killed in war." His self-ascription is an expression of his belief *and also* expresses the fact that he takes the belief to be warranted. If we ask him: "why do you believe this?" he will respond: "because I received a telegram informing me about it." This is a perfectly reasonable way to understand the idea of "expressing your endorsement of a mental state." But it is not how I will be using it. My usage is meant to require something weaker. "To express your endorsement of a mental state" merely entails thinking that it is justified to hold the mental state, even if evidence that one can provide is less than adequate. To illustrate this we might compare the father I just discussed with a father who claims: "I have overwhelming evidence that my daughter was killed in war, nonetheless I still believe that she is alive." This person cannot articulate what warrants his belief. And he can even acknowledge that he does not have good warrants. But because he is committed to this belief, because he thinks that it is true, he believes that there are warrants that justify it (even if he cannot provide such warrants here and now) and even if he can acknowledge that it seems really hard to find such warrants. Thus, when I say that a person expresses her endorsement of the mental state, all I mean is that the person takes the mental state to be justified, even if there is very little evidence that such justification can be provided.

The following tree provides a graphic representation of the relationship between the

different types of self-knowledge discussed so far.

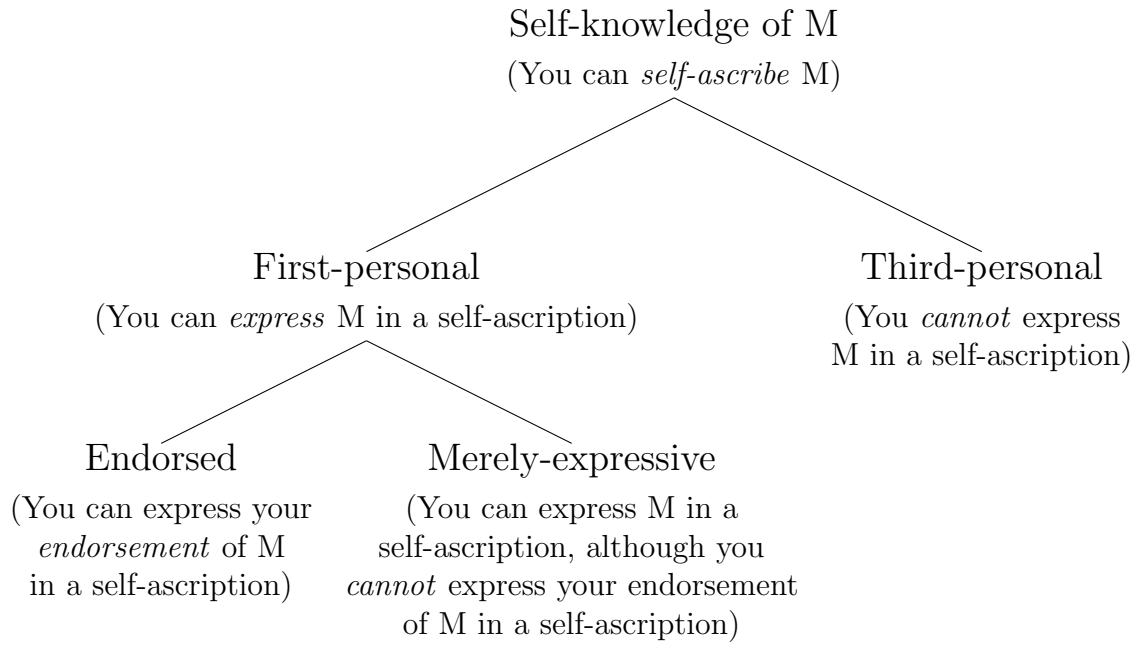

Figure IV.1: A Taxonomy of Self-Knowledge

In the rest of this chapter I will focus on discussing the ethical significance of endorsed self-knowledge. In the next chapter I will discuss why it is also ethically significant to have merely-expressive self-knowledge.

IV.3.2 Two Types of Belief Self-Ascriptions

Establishing an internal connection between *third*-personal self-examination, *third*-personal self-knowledge and *third*-personal grounds, on the one hand, and *first*-personal self-examination, *first*-personal self-knowledge and *expressing* your mental state in a self-ascription, on the other, will allow me to bring out some important features that distinguish endorsed self-knowledge from third-personal self-knowledge.

I'd like to start by discussing the case of belief. When you assert: "I believe P," you are reporting that you hold a belief, namely, P. There are, however, at least two ways in which you can relate to this report. "I believe P" can be a mere report, a report which you make

as a mere witness of yourself. In this report you are merely reporting that you have a belief but your report does not express your endorsement of this belief. I will refer to this type of report as a "witness report." Arguably, not all of your self-ascriptions are witness reports. You can also report on your belief in such a way that this report is also an expression of your endorsement of P. I will refer to this type of report as an "endorsed report."

Suppose that you are on the wake of an election that has important repercussions for your life. In the day of the election you find yourself waking up eagerly and in a very good mood (this is not the usual way in which you wake up). You go early to the voting booth and throughout the day appear cheerful and optimistic. Recognizing all of this behavior might lead you to assert: "I believe that we will win this election." I am imagining that this report corresponds to what I have labeled a "witness report." Your observations of your behavior lead you to ascribe this belief to yourself, but your self-ascription does not express your own endorsement of this belief (as a matter of fact we can suppose that you have not consciously come to a definite position on this matter or even that you consciously deny believing such a thing). Take now a second type of self-ascription. You've been following the news reports and the poll trends of your candidate and you have noticed that, although your candidate has been ahead in the polls, the gap has been closing. There has also been a recent surge of negative media reports about her. This evidence leads you to assert: "I believe that we will lose this election." As I am imagining it, this self-ascription is an endorsed report: it is not merely informing someone that you have a belief; it also expresses your endorsement of this belief.

In our example, the witness report was justified by certain evidence about your own behavior (for instance, that you were cheerful and optimistic). Your endorsed report, by contrast, is not grounded in observations about your behavior; it is grounded in observations about the world (in this case, about the polls and the media reports about the candidate). A witness report is justified by evidence about you to which a third person could appeal

to justify attributing this mental state to you. By contrast in an endorsed report, which expresses your endorsement of the belief P, the evidence that will lead me to utter this report will be internally connected with my endorsement of P, it is evidence about the world (not about me).

This difference comes out clearly in the way in which the following two questions are answered by someone making one or the other report:

1. "How do you know that you believe P?"

2. "What justifies P?"

In the witness report you are expected to respond to the first question by providing evidence about your behavior that warrants attributing P to you.[20] This evidence makes reference to you and not, or only incidentally, to the plausibility of P. In fact, when you provide a witness report you might have no clue about what grounds P and this does not impugn the report's epistemic credentials. In a witness report, then, you need evidence to substantiate attributing a belief to you but you are not required to understand what justifies this belief. When you are making a witness report, you might not have an answer to the second question.

Things look different in the case of an endorsed report. In this case you are expected to have an answer to the second question; you are expected to provide evidence which shows that it is reasonable to hold P.[21] However, when you make an endorsed report, the first question has an odd ring to it. The fact that you believe P seems to follow immediately from the fact that you come to see P as true, that you endorse P. Anyone who speaks your language understands that your endorsement of P warrants self-ascribing this belief. Because of this the question: "How do you know that you believe P?" sounds strange. One might

20. Saying that you are "expected to provide evidence" does not entail that you will need to have this evidence at hand nor that such evidence needs to be particularly solid. As epistemically responsible agents, we are expected to have good evidence to back up our claims, but this expectation is often not met.

21. Once again, saying that you are "expected to have an answer" does not entail you actually have a ready answer to it or that you are capable of fully justifying such an answer. It is just to say that if your answer was epistemically responsible, you would be able to do this.

respond by saying "I know I believe P because I think that P is true." This answer, however, is not much better than the tautological answer "I know that 'I believe P' because I believe P."[22]

The difference between these two types of reports shows up in the nature of accountability that you have about it. Let's start with a witness report. A witness report is a self-ascription of a belief which does not express your endorsement of it. This means that in a witness report you are only reporting that you hold a belief but you are doing so by bracketing or silencing your own endorsement of this belief. That you step back from your endorsement of this belief entails that in your self-ascription you are not making yourself accountable for your belief (as we just said, you might not even know why you hold this belief, all that you know is that there is compelling evidence which warrants attributing this belief to you).

Things are quite different in the case of an endorsed report. As I defined it, your endorsed report is a self-ascription of P which is also an expression of your endorsement of P. Because your endorsement of P infuses your endorsed report, in asserting "I believe P" you are also making yourself accountable for the belief that you hold. Insofar as this utterance expresses your endorsement, you are expected to have evidence to defend P and to change your allegiance to P if you find countervailing evidence that refutes it. But you are not expected to know anything about your behavior, or to make any inferences from it, to make an endorsed report. All you need, to make an endorsed report about P, is to endorse P.

This difference comes out clearly in what it takes to make an epistemically responsible report in each case. Making a witness report in an epistemically responsible way entails being in a position to provide convincing evidence for the self-ascription of P, for the fact that you hold it. Making an endorsed report in an epistemically responsible way entails being in a position to defend the belief, to provide convincing evidence for the truth of P.

22. An alleged tautological answer, however, that is informative. It serves to indicate that your report is an endorsed report and not a witness report.

IV.3.3 Judgment-Sensitive Mental States

What I have just said about the case of belief can be generalized to the case of any judgment-sensitive mental state. Remember that a judgment-sensitive mental state is one that could be formed and is meant to be transformed by the subject's own judgment about whether it is merited to hold the mental state.

In an earlier chapter (II.2.4) I said that I am using the expression "judgment-sensitive" to characterize the class of states of mind that *ought* to be responsive to our judgments even if, as a matter of fact, they are not actually responsive to such judgments. Thus, when I say that a mental state is "judgment-sensitive" I do not only mean that such mental state has been actually formed by a judgment assessing whether one should hold it, nor that it is, in fact, sensitive to it. It only implies that it is the type of state that *could* be formed and *should* be transformed by such a judgment.

Beliefs are judgment-sensitive. But so are intentions, many emotions, as well as a good number of desires. Judging or deliberating about the content of an emotion or a desire is sometimes all that it takes for one to develop this emotion or to have this desire. It is because I judge the reaction of my partner to be inconsiderate and callous that I get angry at her and want her to apologize.

Some readers might have worries about lumping together beliefs, intentions, desires and emotions given that there are significant differences between them. In saying that all of these mental states are judgment-sensitive, I am not attempting to suggest that they can all be reduced to beliefs or that they can all be reduced to different forms of judgment. All I am suggesting is that it is essential to these mental states that they are responsive to the word, that there are rational warrants on which these mental states are supposed to depend.

Emotions, they might argue, seem particularly prone to be misaligned with a person's judgments about their warrants, so much so that it is a mistake to think that they are judgment-sensitive. This worry is unfounded. Even if emotions are more prone to be mis-

aligned with a person's judgments than beliefs, it is still the case that they are, quite often, responsive to these judgments.

The idea that the facts of the matter warrant an emotion is widely accepted.[23] John McDowell has defended that emotions are "merited responses" to the situation.[24] Justin D'arms and Daniel Jacobson, for their part, have used the notion of "fittingness"[25] to describe this fact, suggesting that "the fittingness of an emotion is like the truth of a belief."[26]

Emotions are meant to be aligned with a person's judgments. In fact, when emotions are so aligned, that is to say, when they are fitting, this is usually a sign of mental health.[27] This is not to deny that emotions (and intentions and desires) might have important structural differences with beliefs, it is just to say that being judgment-sensitive is not one of these differences.

I will show below that there is a close connection between a witness report and third-personal self-knowledge on the one hand, and an endorsed report and first-personal self-knowledge, on the other. In particular, I will show that third-personal self-knowledge allows you to make, at most, witness reports, and that only if you have first-personal self-knowledge can you make an endorsed report. Before doing this, however, I would like to elucidate a bit further the distinctive features of these two types of reports.

IV.3.4 Introducing (and Motivating) a Heuristic and a Notation

In this section I would like to make a couple of remarks that will justify a certain heuristic to think about third-personal self-knowledge and a notation that will make perspicuous some

23. Different kinds of defenses of this thesis can be found in: Sousa 1987; Sherman 1991, 2007; Solomon 1999, 2003, 2004; McDowell 1998a; D'Arms and Jacobson 2000a,b; Nussbaum 2001; Goldie 2000, 2010.

24. McDowell 1998a.

25. D'Arms and Jacobson 2000b.

26. D'Arms and Jacobson 2000b, p. 72.

27. I will provide some substantive arguments defending that there is a tight connection between mental health and fittingness in the next chapter.

key features of self-ascriptions that will help me to talk about the topic.

The first of these has to do with a particular way of thinking about third-personal self-knowledge which sheds some light on its nature. My suggestion is that we think of third-personal self-knowledge as "the knowledge that I have of a person who happens to be me." Allow me to explain. I said that you have third-personal self-knowledge of a mental state M if your capacity to self-ascribe M allows you merely to report M but not to express M in a self-ascription. Given what I argued in the previous section, this means that, in third-personal self-knowledge, my self-knowledge is supposed to be explicitly justified by grounds which are available, at least potentially, to a third person.

If we think about this method to acquire knowledge more abstractly, however, we come to see that, as a method to come to know a mental state, it is a method to acquire knowledge which grounds the knowledge which any subject could come to have of the mental states of another subject. In particular, it is a way of knowing which grounds the knowledge that you, yourself, can come to have of the mental states of another subject. Thus, in knowing your mental states third-personally, you are coming to know your mental states in the way in which you come to know the mental states of a third person (a third person who happens to be you).

Coming to realize this provides a good heuristic to think about the nature of third-personal self-knowledge. We have clearer intuitions about how we know the mental states of a third person than we do about how we come to know our mental states third-personally. When posed with a question about third-personal self-knowledge, it is helpful for us to first reflect on how this question is answered when we come to know the mental state of another person and then to transfer our findings to the case of ourselves, i.e. to the case of third-personal self-knowledge.

I mentioned earlier that a witness report is a self-ascription of a mental state which does not express such mental state nor your endorsement of it; in a witness report you are only

reporting that you hold a mental state, but you are doing so by bracketing or silencing your own endorsement of such mental state. Thus, a witness report is a report where the 'I' who is speaks (and who reports "I hold mental state M") is at a distance from the 'I' who holds mental state M. In an endorsed report, by contrast, the report "I hold M" expresses the speaker's endorsement of M. And this means that, in an endorsed report, there is no distance between the 'I' who speaks and the 'I' who believes.

All of this should serve to see that the locution "I hold M" is ambiguous: it can be interpreted either as a witness report or as an endorsed report. The remarks of the previous paragraph bring out that the ambiguity in this locution trades on an ambiguity with respect to the referent of the first person pronoun. As I just said, in the witness report, the 'I' who speaks (and who self-ascribes M) is at a distance from the 'I' who holds this mental state. When you utter a witness report, "I hold M," you are attributing a mental state M to the person who happens to be you. Here, "I" is referring to a person whose mental state can be known, at least potentially, by a third person. By contrast, in the endorsed report "I hold M," the 'I' is a self-referring expression that refers to the speaker of this endorsed report. One could capture this by saying that, in an endorsed report, the speaker in me, the part of me that is speaking at the moment, holds M, while in a witness report, the speaker in me, the part of me that is speaking at the moment, is neutral as to whether she holds M or not.

These remarks motivate the following notation. I will keep using the locution "I hold M" in the way that it is used in common language, that is to say, as ambiguous between a witness or an endorsed report. To indicate that a report is an endorsed report, I will use the locution "I *qua* speaker, hold M." To indicate that a report is a witness report, I will use the locution "The person who happens to be me holds M" or "I, *qua* witness, hold P."

I mentioned above that the distinction between first- and third-personal self-knowledge tracks a distinction between two different ways of being conscious of a mental state (IV.1.1). We can appeal to this distinction to come up with a different locution to disambiguate

between the two senses of the assertion "I hold M." I will also use the locution "I conscious*ly* hold M" to indicate that the assertion is an endorsed report and the locution "I am conscious *of* the fact that I hold M" to indicate that the assertion is a witness report.

IV.3.5 The Ways in Which Endorsed Self-Knowledge Is "First-Personal"

Endorsed self-knowledge is a species of first-personal self-knowledge. As such, the person with endorsed self-knowledge can express a mental state M in a self-ascription. This already entitles one to call this way of self-knowledge "first-personal" because only the subject of M can express M (IV.1.1). Endorsed self-knowledge, however, is first-personal in a deeper way. When a person has endorsed self-knowledge, she makes it the case that she holds M by judging that M is merited. Once again, this is a capacity that is only available to the subject of a mental state. Endorsed self-knowledge is first-personal because only the subject of a mental state has the authority to make up her mind.[28]

IV.3.6 Endorsed Self-Knowledge and Epistemology

While the distinction that I am drawing between first- and third-personal self-knowledge is indebted to David Finkelstein's expressivist account, the distinction between endorsed self-knowledge and merely-expressive self-knowledge is indebted to the agential theories of self-knowledge defended by Richard Moran, Victoria McGeer and Matt Boyle.[29] The work of these authors is meant to defend a particular theory of first-person authority. As with Finkelstein's account, I am retooling some of their insights for a different purpose and within a

28. It is certainly true that I can judge that P, "it is raining outside," and know, in virtue of my judgment, that the person sitting next to me in the bus believes that P. But I know this, not because my judging that P makes it the case that my neighbor believes that P, but because I attribute to my neighbor the same capacity to make up her mind that I myself have deployed. I am assuming that my neighbor believes that P *because* she herself has judged that P.

29. Richard Moran 2001; McGeer 1996, 2007; Boyle 2009, 2011.

different project. I have found that these accounts provide insightful ways to understand how we are related to the mental states over which we take ourselves to have immediate, private and authoritative self-knowledge. But I use these insights to conceptualize a certain way in which we can relate with our self-knowledge, regardless of how this knowledge is ultimately acquired. As a consequence, and as I have already mentioned, my account is meant to be neutral with respect to the theories of self-knowledge proposed by epistemologists.

IV.4 Endorsed Self-Knowledge Is Ineliminable

The conceptual work we've done in this chapter allows us to examine the particular role that first-personal self-knowledge plays in human life and, thereby, to justify why it is important for the ethical pilgrim to know her mental states first-personally. It has been my experience that, upon hearing this thesis, namely, that first-personal self-knowledge is ethically significant, readers are divided in three groups. The first, a small group, thinks that this position is obviously and necessarily true. The second, the biggest group of the three, is composed by people who find the thesis plausible but are curious to understand what it is that makes it true. Finally, there is a third group, also constituted by few people, which does not think that it makes any difference, from an ethical point of view, whether one knows oneself first- or third-personally. In showing why first-personal self-knowledge is ethically significant I will show the third group that they are mistaken and will provide the second group with what they want to understand. In doing so, it will also show the first group that the argument for this claim is neither obvious nor conceptually necessary.

IV.4.1 A Mental Experiment

My discussion of Wilson in the previous chapter appealed to the kinds of resources for self-examination and self-knowledge that are currently available to human beings. But to

properly understand why first-personal self-knowledge is significant for ethical development, why third-personal self-knowledge cannot be all that the ethical aspirant shall aspire to acquire, it will be necessary to strengthen as much as possible Wilson's position. And to do so we should pursue a thought experiment in which third-personal self-knowledge has all the virtues that Wilson attributes to it. So place yourself in a future where technology and neuroscience have advanced so much that you can buy an Inner Self-Detector, an extremely reliable app that can tell you about your mental states. The advertisement in the box states: "No more opaque soul-searching! The Inner Self-Detector provides you with the most reliable and accurate access to your mental states. If you want to know what you desire, intend, believe or feel, just ask the Inner Self-Detector." The advertising of the product claims that empirical tests have firmly established that this app is more reliable than introspection. I will try to make sense of the idea that a person can live her life guided only by third-personal self-knowledge by imagining a scenario where a particular person, Chloe, heeds Wilson's advice and attempts to rely on the Inner Self-Detector to know her mental states.

Wilson invites us to relinquish first-personal self-examination and to foster third-personal self-examination instead. There are moments in his book where it appears that Wilson seems to believe that it is coherent to think that we can rely only on third-personal self-knowledge of our mental states, that it is possible for creatures like us to completely relinquish first-personal self-knowledge. This, as I will now show, cannot be right; no matter how much we examine ourselves third-personally, no matter how much we rely on an Inner Self-Detector, first-personal self-knowledge will still be ubiquitous and ineliminable (IV.6). I will then show that first-personal self-knowledge is also indispensable (IV.5) and that it plays an important role in the pilgrim's attempts to develop virtue. The ethical pilgrim, I will argue, should not be content with the kind of self-knowledge that the Inner Self-Detector provides. Even if the information provided by it is valuable for her, she should aspire to eventually come to know her mental states first-personally.

In this chapter I will focus on showing why this is the case for one particular type of first-personal self-knowledge, endorsed first-personal self-knowledge. I will devote the next chapter to bring out the role that another type of first-personal self-knowledge, merely-expressive self-knowledge, plays in the development of virtue.

Before doing so, it is appropriate to mention a caveat. I will criticize the idea that the ethical pilgrim should be content with knowing about her mental states only through the use of the Inner Self-Detector. But it is important to highlight that this app is a very valuable addition to her "ethical development toolkit." As I said in the conclusion to the previous chapter, we should be less confident about the reliability of our self-examination and more humble about our capacity to provide accurate reports about ourselves. An Inner Self-Detector would improve the accuracy of our self-investigations and, in doing so, would help us triangulate information about ourselves that can help us improve our knowledge of our mental states. And nothing of what I will argue below is meant to put this into question. My disagreement with Wilson is on the thesis that this kind of self-knowledge will be sufficient for the ethical aspirant.

IV.4.2 Higher-Order Mental States

To start exploring our thought experiment, let's focus on an everyday moment in Chloe's life. Mark, Chloe's friend, calls and asks her whether she would like to hang out with him in the afternoon. Chloe responds: "Give me a second while I ask my Inner Self-Detector whether I'd like to do this." She opens the app in her cell phone and asks: "Do I want to hang out with Mark?" A robotic voice responds: "Yes you do." Upon hearing this Chloe tells Mark: "Yes, I do." Mark responds: "What would you like to do?" Once again Chloe asks the Inner Self-Detector. Upon hearing the app's response she tells Mark: "I want to go to the movies." To show that endorsed self-knowledge is ineliminable and pervasive I'd like us to reflect on the relationship that Chloe has to her third-personal self-ascription Q:

"I want to go to the movies" (or, to be more precise, "The person who happens to be me wants to go to the movies"). What I want to argue is that this witness report entitles Chloe to know, with endorsed self-knowledge, her belief R: "I, *qua* speaker, believe Q."

A *witness* report is epistemically responsible when the person is able to provide convincing evidence that warrants self-ascribing it. The fact that the Inner Self-Detector is reliable suggests that the Detector's testimony provides Chloe with such warrants. And because of these warrants, her witness report will be knowledgeable, it will be an instance of third-personal self-knowledge.

An *endorsed* report is epistemically responsible when the person is able to provide warrants for the merits of holding the mental state (and not, like in the witness report, warrants to self-ascribe this mental state to her). This means that Chloe can make the endorsed report R ("I believe Q") in an epistemically responsible way. This is so because, as I just said, Chloe has good warrants, i.e. the Detector's testimony, to hold the belief Q. Chloe's capacity to assert an endorsed report R entails that she has first-personal self-knowledge of R.

This, of course, generalizes to any mental state. Chloe's capacity to make a witness report "I hold M" in a responsible way entails, *eo ipso,* her capacity to make the endorsed report "I, *qua* speaker, believe 'I hold M'" in an epistemically responsible way. This is so because she has warrants to believe "I hold M", and these warrants entitle her to assert "I *qua* speaker believe that 'I hold M'." Thus, Chloe's third-personal self-knowledge of "I hold M" provides her with first-personal self-knowledge of "I believe 'I hold M'."

Here is an alternative way to make this point. When Chloe *asserts* Q she is in a "privileged" relationship with Q; when she asserts Q, she is in the position to express her endorsement to "I believe Q." Her capacity to assert "I want to go to the movies" presupposes her capacity to assert R: "I, *qua* speaker, believe 'I want to go to the movies'."

This entails that every nugget of third-personal self-knowledge entitles Chloe to a nugget

of first-personal self-knowledge; every (truthful) assertion where she self-ascribes a mental state M entitles her to assert the endorsed report "I believe I hold M." And because this holds regardless of whether the first assertion is a witness report or an endorsed report it follows that the amount of beliefs about which Chloe can know first-personally are actually pervasive.[30]

This argument is unlikely to impress Wilson. He could grant that first-personal self-knowledge is pervasive in the way I am suggesting but reply that his suggestion to examine (and know) ourselves third-personally is circumscribed to first-*order* mental states. The argument I provided does not show that first-personal self-knowledge of our first-*order* mental states is pervasive. It only shows that first-personal self-knowledge *of higher-order mental states* is pervasive.

IV.4.3 First-Order Mental States About Which You Make up Your Mind

To show that first-personal self-knowledge of first-*order* mental states is ubiquitous and ineliminable it will be helpful to follow the conversation between Chloe and Mark for a bit longer. Upon hearing Chloe say that she wants to go to the movies, Mark adds: "Cool. Would you like to watch *Chi-Raq*? It's playing at the Hyde Park Theater." Chloe asks the Inner Self-Detector: "Do I want to watch *Chi-Raq*?" The robotic voice responds: "You don't know." The detector responds this because Chloe has never heard of this movie; she does not know what this movie is about, who has directed it or who acts in it. Upon hearing the Detector's response Chloe thinks to herself: "Of course I don't know whether I want to go;

30. In fact, for every mental state M that is a person knows with third-personal self-knowledge there is an infinite number of mental states that she knows first-personally. Her second-order assertion R, "I, *qua* speaker, believe 'I hold M'" entitles her to make the third-order assertion S: "I, *qua* speaker, believe R." And this third-order assertion allows her to make a fourth-order assertion T: "I, *qua* speaker, believe R" and so forth. Each assertion she makes, puts her in a position to make a higher order assertion about the fact that she believes what she asserts.

I have never heard of this movie! I need to look it up." She opens her phone's browser and realizes that Spike Lee directs it. She dislikes Spike Lee's movies so she tells Mark: "I don't like Spike Lee's movies; I don't want to watch *Chi-Raq*!"[31]

Chloe is here considering features of the world, in this particular case features of a particular movie, which warrant and justify her desire not to watch *Chi-Raq*. When she tells Mark "I do not want to watch *Chi-Raq*!" her remark is an endorsed report. She does not merely know that she has a desire not to watch it, she endorses this desire in light of the considerations of the world that merit such desire. Someone sympathetic to Wilson might complain that, here, Chloe is not following his advice; before telling Mark whether she wants to see the movie or not she should consult the Inner Self-Detector. This will not make too much of a difference to what I have just said. Let's allow that Chloe decides to ask the Inner Self-Detector whether she wants to see *Chi-Raq*. The app's response will provide Chloe with third-personal self-knowledge. As a result, she will be able to make a witness report based on the app's response. But the fact of the matter is that Chloe will still have first-personal self-knowledge of her desire not to go to see this particular movie. Even if Chloe can make a witness report about her desire, she is also in a position to make an endorsed report. Unless she has a pathologically bad memory, Chloe will be able to remember the considerations that merit her desire, she will be able to identify with these considerations and, as a consequence, she will be capable to make a self-ascription that expresses her endorsement of this desire, i.e. she will have first-personal self-knowledge.

The main lesson that we can draw from this is that, whenever a person has to make up her mind, this person will have first-personal self-knowledge. Thus, there are at least as many first-order mental states about which we have first-personal self-knowledge as mental

31. I am portraying the Inner Self-Detector, here, as an app that can merely report about a person's mind here and now, an app that can report on the desires, intentions, beliefs, or emotions that a person currently holds. If a person has not made up her mind the Inner Self-Detector will report this.
Later in the chapter I will consider a more powerful Inner Self-Detector, one that incorporates and computes information from the world that allows it to anticipate what will be the mental states about which Chloe has not yet made up her mind.

states about which we have to make up our minds.

Realizing this allows us to go back and identify a number of mental states in our vignette about which Chloe has first-personal self-knowledge. Her intention to look up the movie is of a first-personal kind; it was warranted by her need to look up information about *Chi-Raq*, information that she required to determine whether to watch it or not. Chloe is in a position to report "I, *qua* speaker, intend to look up information about *Chi-Raq*." Similarly, her knowledge that she has not heard about *Chi-Raq* is also first-personal. It is information that was not provided by the Inner Self-Detector. Finally, Chloe also has first-personal self-knowledge of each of her decisions to ask the app to report on a particular mental state. Each of these decisions is warranted by her conviction that it is important to rely on third-personal self-knowledge like the one provided by the app.

IV.4.4 Mental States About Which You Have Already Made up Your Mind

A follower of Wilson might still be unimpressed. She might point out, following an influential paper by David Velleman and Nishi Shah, that "[t]he question 'Do I believe that P' can mean either 'Do I already believe that P (that is, antecedently to considering this question)?' or 'Do I now believe that P (that is, now that I am answering [whether P is true])?"[32] Wilson might concede that when you are making up your mind you have first-personal self-knowledge. But he might respond that his proposal applies only to cases where the question is understood in the first sense, not in the second, cases in which we have already made up our minds.

As I will show below, however, in many of these cases we also have first-personal self-knowledge. To see why, let us consider a few more lines in the conversation between Chloe and Mark. Upon hearing that Chloe is not interested in watching *Chi-Raq*, Mark might add: "We could perhaps watch The Big Short. I don't really like financial movies, but the reviews

32. Shah and Velleman 2005, p. 506.

I've read are quite promising." Chloe, who has heard and read a number of things about this second movie, might ask the Inner Self-Detector: "Do I want to watch *The Big Short*?" The robotic voice will respond: "No, you do not." There are two possibilities here. Either the warrants of Chloe's desire not to watch this movie are available to her or they are not. I'd like to conclude this section by focusing on the first possibility and to turn to the second in the next section.

If Chloe can identify (and stand by) her warrants to desire not to go to The Big Short then she would have first-personal self-knowledge of this desire. The Inner Self-Detector would be providing her with third-personal self-knowledge. But her third-personal self-knowledge would be an addition to her first-personal self-knowledge. It would merely confirm it. If Chloe can recognize why she does not want to go to The Big Short, namely, because she cannot stand seeing Brad Pitt, then she will be in a position to make the endorsed report: "I don't want to go to *The Big Short*."

This result shows that first-personal self-knowledge is very pervasive and ubiquitous. It shows that we have first-personal self-knowledge, not only of mental states about which we are making up our minds here and now, but also of mental states about which we made up our minds in the past, provided that we remember (and still stand by) the warrants that justify them.

IV.5 The Dialectic Between Detecting and Constituting a Mental State

In the previous section I argued that endorsed self-knowledge is ubiquitous and ineliminable. What I will now argue is, not merely that we cannot dispense with first-personal self-knowledge, but that we would be ill-served if we were to do so.

IV.5.1 Detecting and Constituting; Discussing and Convincing

To see this, let's consider the second alternative mentioned above: the Self-Detector tells Chloe that she does not desire to go to the movie and Chloe does not know why she has this desire (that is to say, her warrants are not available to her consciousness). This is a case in which we can assume that Chloe is speaking only with third-personal self-knowledge. Consider, now, one possible way in which the conversation could have continued. Upon learning that Chloe does not want to go to The Big Short, Mark is surprised and replies: "Really? You don't want to watch The Big Short? You love movies about Wall-Street! It has great reviews and, besides, Christian Bale acts on it. You love that dude. You should reconsider it!" If Chloe's report is third-personal, then it is not clear that she is in a position to consciously reconsider her view. In fact, as I will now argue, if Chloe is disconnected from the reasons for her desire, it appears that Mark and Chloe have no way of arguing with one another about her desire.

The reader might think that Chloe could have access to her reasons by asking the Inner Self-Detector. But this will not allow Chloe to properly discuss this issue with Mark in the way that we usually discuss it with our friends. If we want to keep thinking that all that Chloe has to go by is third-personal self-knowledge, we would need to imagine that, even if she found out what her reasons are for not wanting to go to the movie, she would need to be disconnected from these reasons. Her endorsement of her reasons would have to be kept at bay. If Chloe can, at most, make witness reports, it is hard to see how Mark and Chloe can have a discussion about the merits of going to The Big Short. Chloe will be merely reporting on the desire of the person who happens to be her. But she will be acting as a witness about Chloe who does not endorse Chloe's desire not to go to the movie. When Mark speaks with Chloe, *qua* witness, it is as though he is speaking with someone else about Chloe (someone else who just happens to be Chloe); she is like a mere spokeswoman for Chloe's mental life, a spokesman who can only report on it but who can play no role in determining

or constituting such mental life. The reasons that Mark raises do not come in contact with the reasons that Chloe holds *qua* speaker because they are mediated by a witness who does not endorse Chloe's reasons. The witness cannot defend Chloe's desire. As a consequence it is very difficult to see how Mark could convince Chloe.[33] If a person can only know her mental state third-personally, then she is incapable of engaging in a conversation which has the potential to transform this mental state.[34]

This suggests a central reason why it is ethically significant to have first-personal self-knowledge. As I will conclude at the end of this section, to be able to articulate a desire in speech does not merely involve being able to report it. It also involves being able to defend it when its warrants are criticized and to change it when its warrants are undermined. As we will see, the person with third-personal self-knowledge can only report a mental state but she cannot discuss it with others in a way that opens this mental state to revision. Her inability to do this shows the limitations of third-personal self-knowledge.

A judgment-sensitive mental state is not just a lifeless piece of furniture lying around your mind. A judgment-sensitive mental state is as much about you as it is about the world around you. Although a judgment-sensitive mental state describes something about your mind it is also meant to be responsive to the world. I believe that my candidate will win because of what the polls suggest and I am angry at you because you slighted me. If things are going well, the question: "What do I believe/intend/desire/feel?" is not independent from the question "What *is warranted* to believe/intend/desire/feel?" Our capacity to make judgments about the world is intimately tied to the mental states that ensue as a result. And the fact that we can reflect and have language to think about these mental states entails

33. Later in this section I will consider one possible way in which this might happen and will indicate what is problematic about it.

34. To be more precise, what I mean to say is that "she is incapable of engaging in a conversation which has the potential to transform this mental state *in a non accidental way*," that is to say, to transform it because the mental state responds to the conversation and not because the conversation brings about certain happenings that end up transforming it.

that we recognize this fact and are explicitly aware of our ability to constitute these mental states in light of these judgments.

Velleman and Shah have challenged this view, arguing that: "[O]ne cannot engage in reasoning aimed at answering the question whether P if one wants to find out what one already believes, because such reasoning would contaminate the result by possibly altering the state that one is trying to assay."[35] Now, if the question is a question about what you believed, in the past, I agree with Velleman and Shah that the question whether P can contaminate our answer. But if one is meant to report on what one believes, here and now, then one's report is not expected to be a mere report of some inert content lying around in your mind, independent of the way in which I am, here and now, perceiving the world. Because our belief is *about* the world and is meant to be responsive *to* the world, the question of what I believe cannot be independent of the question of what is warranted to believe. What Velleman and Shah call the "possibility of contamination" of the mental state should rather be seen as an "opportunity for the perfection" of such mental state. That the question "what do you believe?" opens up the question "what should you believe?" is actually an opportunity for the mental state to perfect itself, to respond to the world adequately.

Allow me to generalize this to other judgment-sensitive mental states. Typically, our report of a mental state that we currently hold is pregnant with the possibility of revising and updating it. A judgment-sensitive mental state is not an object lying in our minds ready to be witnessed. It is active, responsive to the world, in dynamic relationship with it. As I will argue more fully in the next section, when things are working well, the judgment-sensitive mental state of the person will be always open to be updated and transformed in response to what the person conceives to be warranted. Thus, what Velleman and Shah call the possibility of contamination of our mental states is, not merely at the heart of their nature, but actually exhibits their perfection.

35. Shah and Velleman 2005, p. 507.

IV.5.2 "Do You Want to Marry Me?"

I'd like to follow up Mark and Chloe through their friendship and subsequent romance. A couple of years have passed and Chloe learns that Mark is planning to propose to her. Chloe, following Wilson's advice, decides to ask the Inner Self-Detector whether she wants to marry Mark. Reflecting on this scenario will bring out to the fore some of the issues that are problematic with a contemplative conception of self-knowledge like Wilson's, where self-knowledge is conceived in terms of the mere acquisition of information.

What I want to argue requires stipulating that Chloe's situation is such that the answer to this question is not entirely clear to her. With Mark, as one says, one thing led to another and she now looks back and realizes that she has been dating him for almost two years. Time has flown and she has been quite happy with him. But marrying him? She had not really thought about it and she is not clear about what the thinks about this. Is Mark the person with whom she would like to spend the rest of her life? There are things about him that are really appealing and others which are not; he is great with kids and not very demanding, but he is often stingy and is pretty disorganized. If Chloe asks the Inner Self-Detector the answer will be: "you don't know." In this case, however, she does not know, not because she lacks information (like with *Chi-Raq*). Chloe does not know whether to marry Mark, not because she lacks information but because she lacks clarity on how to answer the question.

The question: "do I want to marry Mark?" is, for Chloe, not just the question of whether to marry Mark. It is a question that is entangled with more general questions such as: "What is the type of man that I want to marry?", "What does marriage mean for me?" and "Do I want to marry at all?" Responding the first question, the question about whether to marry Mark, requires her to be in a position to have at least a rough answer to the other questions. This scenario helps us to see some of the limitations of the Inner Self-Detector. Reporting on our mental states is not just a matter of reporting on settled attitudes. Knowing what her judgment-sensitive mental states are and judging about the content of these mental

states are not independent activities; they mutually entail one another. Shah and Velleman suggest that raising any of these questions will contaminate the response about what Chloe thinks. But the locution "contaminate" does not do justice to the dialectic at play here. The articulation of the response does not contaminate the response because the response is not meant to be a formulation in words of an already settled mental state. The articulation of the response is an opportunity to get clearer on what the question is asking, the different alternatives to respond to it and one's definite take on it. Chloe's struggle to respond to this question is a struggle to determine the shape that her life is meant to take. It is not merely a response about her own internal mental states but also about the kind of life that a human being like her ought to live.

Moreover, articulating responses to these questions is an opportunity to further clarify her mind even when the person has already settled on a response. Thus, even if the Inner Self-Detector could tell Chloe that she does not want to marry Mark, having endorsed self-knowledge would allow Chloe to report on this fact in such a way that her report allows her to reconsider this response. I recently came across an article about the importance of teaching students to write where its author, Stuart Walesh, argued that "writing is one way, and a very effective way, to clarify our thinking."[36] He cites the playwright Edward Albee who wrote "I write to find out what I'm thinking." Albee did not mean, however, that writing helps him to guess what he is thinking. His point is that writing allows him to make up his mind and, thereby, to know what he thinks. And this is true even in the case where the person has already made up her mind. Speaking (or writing) about what one thinks, feels or desires allows one to reconsider it. Endorsed self-knowledge does not merely provide the person with an inert mental state that is lying in her mind, it opens up a process that allows for a re-articulation and reconstitution of such mental state.

Allow me to pause and think for a second on this scenario from the point of view of

36. Walesh 2016.

Mark. Let's ask ourselves what he would feel when Chloe decides to figure out whether she wants to marry him by asking an Inner Self-Detector. Mark is probably not merely hoping that Chloe can report on the fact that she wants to marry him. Most likely he wants her assertion to express her endorsement of her desire of marrying him. He does not want her to repeat what the robotic voice said, but to accept his proposal wholeheartedly. Chloe's witness report: "I want to marry you" reveals to Mark that she is out of touch with her own feelings. And this would make Mark, at the very least, sad. And he would be sad because he wants her, not merely to accept his proposal, but to stand by this proposal. If all that she has to go by is third-personal self-knowledge, Chloe will not be able to do this. Furthermore, if her knowledge that she wants to marry him is only third-personal, then she would not be recognizing why it would be worthwhile to marry Mark. And it is natural to suppose that this is something else that would disappoint Mark. Because Mark wants Chloe to express her endorsement of her decision to marry him because in doing so she reveals that she recognizes why it is merited to marry Mark.[37]

IV.5.3 McGeer and Wilson's Alternative

My general orientation in this chapter, and especially in the last two sections, is highly indebted to the work of epistemologists like Moran, McGeer and Boyle.[38] These authors put forth what is often referred to as "agentialist"[39] or "rationalist"[40] accounts of self-knowledge.

Anyone familiar with this tradition will recognize their influence in what I have been writing. In what follows I want to bring out that there is an unargued premise in their

37. None of this is to suggest that marrying someone is a merely rational decision based on the warrants of the potential spouse. My suggestion here is only that when I am marrying someone I would like my fiancé to be fully committed to this. Arguably, this requires her to have endorsed self-knowledge. This, however, is not available to the person who is only being guided by third-personal self-knowledge.

38. Richard Moran 2001; McGeer 1996, 2007; Boyle 2009, 2011.

39. Gertler 2011.

40. Cassam 2014.

account. These authors identify our 'capacity to speak' with our 'capacity to reason,' but they do not provide any argument defending this identification. This unargued premise also haunts my account and requires me to respond to it.

To do so, I will start by considering a theoretical alternative inspired by work in social psychology. This alternative helps us realize that it there are cases where, at least in theory, an organism's rationality can be divorced from her capacity to speak; it is theoretically possible for a creature to have reasonable responses to the world that need not be articulable in speech or mediated by speech. After laying out this possibility I will argue that, although it is a genuine theoretical possibility, it is not instantiated in creatures like us. Human beings use their speech (and ought to use it) not merely to report on their mental states but to make judgments on the world, judgments which serve to constitute their own mental states. This, however, is an empirical point, not a theoretical one.

I mentioned earlier that upon hearing this thesis that first-personal self-knowledge is ethically significant, a first group of interlocutors believes that the position is obviously and necessarily true. However, to the extent that the argument depends on empirical considerations, it would suggest that it is not an argument that can be assumed to be obvious or conceptually necessary.

The scenario that I will construct is inspired by the work of Timothy Wilson and Ap Dijksterhuis. Wilson portrays human beings as divided between two selves: the conscious self and the adaptive unconscious.[41] He writes: "it makes little sense to talk about a single 'self' when we consider that both the adaptive unconscious and the conscious self have regular patterns of responding to the social world."[42] According to Wilson, judgment-sensitive mental states can be attributed to both selves,[43] both selves are relatively independent and they predict different kinds of behavior. About this last point he writes: "The adaptive

41. Timothy D. Wilson 2002, p. 72.

42. Timothy D. Wilson 2002, p. 68.

43. Timothy D. Wilson 2002, p. 125.

unconscious is more likely to influence people's uncontrolled, implicit responses, whereas the [... conscious] self is more likely to influence people's deliberative, explicit responses."[44]

Wilson and his collaborators asked college students to assess how satisfied they were with their dating relationships. Some students had to provide reasons supporting their assessment while others were not meant to analyze their feelings but to respond merely about their gut feelings. The results suggest that "the feelings people report after analyzing reasons are often incorrect, in the sense that they lead to decisions that people later regret, do not predict their later behavior very well, and correspond poorly with the opinion of experts."[45] Experimental results like this suggest that human beings make better judgments when they do not ponder carefully on the reasons that prompt these judgments. Consequently, they lead Wilson to invite us to rely on our gut feelings to make these decisions and not on an explicit reflection on our reasons underlying these judgments.

Wilson mentions that after this study was published a reporter asked him: "So, Dr. Wilson, I gather you are saying that people should never think about why they feel the way they do and should simply act on their first impulses?" Wilson was horrified. In the book he responded to this question by saying: "It is important to distinguish between informed and uninformed gut feelings. We should gather as much information as possible, to allow our adaptive unconscious to make a stable, informed evaluation rather than an ill-informed one. (...) The point is that we should not analyze the information in an overly deliberate, conscious manner, constantly making explicit lists of pluses and minuses. We should let our adaptive unconscious do the job of forming reliable feelings and then trust those feelings, even if we cannot explain them entirely."[46]

Wilson's research suggests that it is better to make decisions unconsciously than consciously. Dijksterhuis' research suggests something similar. In a number of studies, he and

44. Timothy D. Wilson 2002, p. 73.

45. Timothy D. Wilson 2002, p. 170. I discussed this experiment earlier (III.1.2).

46. Timothy D. Wilson 2002, pp. 171–2.

his collaborators asked participants to select a product from a number of items with different characteristics. In a typical version of this study some participants had to make the choice immediately, others where given a fixed time to think about the choice consciously and a third group was asked to make their choice after the same time but while being distracted doing other tasks. The third group was engaging in what Dijksterhuis has called unconscious thought: "a cognitive process in the absence of conscious attention."[47] The results showed that if the decision involved simple choices it was better to engage in conscious thought but if it involved complex choices it was better to allow one's mind to make these decisions through unconscious thought. Because ethical decisions tend to be complex, Dijksterhuis' research would suggest that we should make ethical decisions unconsciously. Dijksterhuis' conclusions are quite similar to Wilson's: "To come to good decisions, what we need is not more self-knowledge. Well, maybe one very simple piece of self-knowledge: To know that when we have a complex decision to make, we do not have to rely on conscious thought. What we need is to encode all information about a decision problem we face and then trust our unconscious to make a good decision."[48] Wilson's and Dijksterhuis' research suggests, then, not merely that human beings are capable of making ethically significant decisions without conscious awareness, but that they *should* try to make these decisions in this way.[49]

This would provide support to the view that it would be better for a human being to know herself third-personally. Although it is not explicitly discussed in Wilson's and Dijksterhuis' studies, it is arguably the case that when the participants reported on their choice, their reports were endorsed reports. An endorsed report wears on its sleeve the possibility of revising the content of what one reports. But given that it is better to make

47. Bos and Dijksterhuis 2012, p. 188. See also, Dijksterhuis and Nordgren 2006; Dijksterhuis, Bos, et al. 2006. It might be relevant to mention explicitly that "unconscious thought" is different from "snap judgments"; unconscious thought requires time whereas snap judgments do not.

48. Bos and Dijksterhuis 2012, p. 189.

49. Further research programs within social psychology would lend further support to this. See for instance, those influenced by Bechara et al. 1997.

decisions unconsciously, this is a possibility that it would be better not to open up. Because an endorsed report opens up the space to demand justifications, Wilson and Dijksterhuis should actually advice people, not merely to make decisions unconsciously but to try and report them third-personally.

IV.5.4 "Do You Want to Marry Me?" (Wilson's Proposal)

Let me return to Chloe and articulate how this research bears on our thought experiment. It is interesting that Wilson also appeals to the example of marriage to exemplify his suggestion. He writes: "Most of us would agree that it would not be wise to marry the first person we are attracted to. If we spend a lot of time with someone and get to know him or her very well, and still have a very positive gut feeling, that is a good sign. (...) The trick is to gather enough information to develop an informed gut feeling and then not analyze that feeling too much. There is a great deal of information we need in order to know whether someone would make a good partner, much of it processed by our adaptive unconscious."[50]

According to Wilson and Dijksterhuis when Chloe has to think whether to marry Mark she should not reflect on this question explicitly. Given that it is an important decision, however, she should spend some time thinking about it. Thus, to make sure that Chloe engages with this question using unconscious thinking frequently, she might prime her unconscious by placing Post-It notes around her house with the question: "Should I marry Mark?" These Post-It notes will activate her unconscious thinking about this question. Sometimes, however, she would find herself reading the question consciously and might be tempted to attempt to think about it consciously. It is for these occasions that she carries around a Sudoku book. Every time that she feels tempted to think about this question consciously, she uses the book to get distracted, ensuring that her engagement with the question is always unconscious. When the day finally comes when Mark kneels, ring in hand, and

50. Timothy D. Wilson 2002, p. 171.

proposes, all that Chloe needs to do is to ask the Inner Self-Detector: "Do I want to marry Mark?"

I said earlier that in addressing the particular question of whether to marry Mark, Chloe had to address other questions such as what is the kind of man she wants to be married to, what marriage means to her and whether she wants to get married. Wilson and Dijksterhuis would argue that the process of unconscious thinking about the first question will also entail unconsciously thinking about the others.

I said above that if Chloe lacks first-personal self-knowledge of her mental state M, it was hard to see how Mark could change Chloe's mind or how they could have a discussion about it. If the research by Wilson and Dijksterhuis is correct, however, there is a way for this to happen. Suppose that Chloe tells Mark: "The Inner Self-Detector tells me I do not want to marry you. I am really sorry." Mark, who is not used to taking "no" for an answer, might reply: "Why? I thought you really loved me!" If Chloe only has third-personal self-knowledge of her desire not to marry him, then she is unlikely to be able to respond to this question off the top of her head and will need to ask the Inner Self-Detector. When the robotic voice responds "Mark is too thrifty and you (unconsciously) think that this will not be good in the long term," Mark might in turn reply: "Is this the main reason? Look, I am certainly thrifty with myself but I do not intend to be thrifty with you. I will allow you to spend as much as you want. You will manage our bank account and all the money I make will go into it." Chloe can see how this might address her concern, and she might respond: "Mmm... I was not aware that you were intending to do this. Let me think about it for a little bit." She turns on the timer on her phone, takes out her Sudoku book and starts doing a puzzle. When the beeper rings, she asks the Inner Self-Detector: "What do I think about Mark's reply? Do I want to marry him given what he has replied?"

Wilson and Dijksterhuis would say that although this is not the way in which we are used have discussions with our friends, it is the way in which we should conduct ourselves. This

is the procedure through which we should learn to change one another's minds. They would also say that Mark's desires to have Chloe consciously endorse or reject his proposal, and his sadness about her not doing so are part of an old fashioned folk psychology that needs to be updated. They would argue that if Mark had read more widely the psychological literature, he would recognize that what he should aspire is to be recognized by Chloe's unconscious self not by her conscious self.

IV.5.5 A Return to McGeer

Having considered this possibility let me return to McGeer. McGeer raises a number of criticisms against inner-sense theories, criticisms that can be retooled to criticize contemplative conceptions of self-knowledge like Wilson's. According to McGeer, a contemplative conception of self-knowledge fails to explain the close connection between acknowledging someone's authority over her own psychological states and treating her as the sort of agent who can be held responsible for what she thinks and does.[51] She writes:

> The agent has a privileged authority in self-ascribing intentional states because it is she who makes it the case that she deserves to be ascribed these states; she has 'maker's knowledge,' not the knowledge of a particularly accurate perceiver or detector.[52]

McGeer ties this responsibility with the person's rationality. She would suggest that a person like Chloe, who operates only with a contemplative conception of self-knowledge, lacks certain "psychological capabilities." In particular, McGeer would argue that she would lack the capacity to think and operate as a rational, responsible and self-directed agent.[53]

51. McGeer 2007, p. 81.

52. McGeer 2007, p. 82.

53. McGeer 2007, pp. 81–2.

Without saying more, these challenges are unfair. If Wilson and Dijksterhuis are correct we can predicate of Chloe all of these psychological capacities that, according to McGeer she does not have. This is so because, Wilson and Dijksterhuis could appeal to the fact that thinking unconsciously leads to better decisions,[54] to argue that Chloe's decisions are not merely better but also more rational. If this is the case, then Chloe's reliance on her unconscious thinking would make her more rational and, therefore, more epistemically responsible as an agent. It is true that this rationality is not mediated by Chloe *qua* speaker. After all, when she is speaking with third-personal self-knowledge the I who speaks about a mental state M is removed from the I who holds M. But because this way to approach her mental states is superior, it would actually suggest that Chloe's reliance on third-personal self-examination and self-knowledge is, in fact, a sign of her superior rationality.

McGeer mentions that a contemplative conception of self-knowledge fails to explain the close connection between acknowledging someone's authority over her own psychological states and treating her as the sort of agent who can be held responsible for what she thinks and does.[55] Once again, this complaint seems misguided about Chloe. I don't see any good reason to think that Chloe is not responsible for what she thinks and does. She is coming up with these decisions on her own; no one is making these decisions for her. Sure, she is not making these decisions consciously. She is not making these decisions *qua* speaker. But even if Chloe is not responsible *qua* speaker, Mark takes her to be responsible *qua* person. When Mark challenges her response he is challenging Chloe's views, not the views of someone else. Mark is assuming that it is Chloe (*qua* person, not *qua* speaker) who has the epistemic agency over her mental states. And when he asks her to reconsider her decision in light of Mark's own response, both Mark and Chloe take it to be the case that this reconsideration is the product of Chloe's own (unconscious) epistemic agency.

54. Bos and Dijksterhuis 2012, p. 182.

55. McGeer 2007, p. 81.

In this section I have argued that there is an internal connection between making judgments about the world and having mental states that reflect these judgments. But this connection holds if Chloe could actually proceed as Wilson and Dijksterhuis suggest. Within their picture there is an internal connection between making (unconscious) judgments about the world and having (unconscious) mental states that track these judgments. In the beginning of this section I tentatively suggested that this internal connection was at the heart of endorsed self-knowledge. I was tentative because it was clear to me that fully arguing for this required showing that our capacity to speak was connected with the dialectic of judging the world and holding mental states that reflected such judgments. If Wilson and Dijksterhuis are right about our power of unconscious judgment this would not hold. The dialectic constitutive interactions between our judgments and our mental states could take place below our conscious awareness. If so, we would not need endorsed self-knowledge to be able to make up our minds in this way.

IV.6 The Importance of Endorsed Self-Knowledge

IV.6.1 A Critical Review of Wilson and Dijksterhuis

Wilson and Dijksterhuis help us to see that it is theoretically possible to have reasonable responses to the world that need not be articulated by us in speech or mediated by it. Their proposals, in fact, suggest that our capacity to speak actually interferes with making proper decisions on complex matters. Although these two social psychologists grant that our capacity to speak is not merely the epiphenomenal whistle of the locomotive of our minds, they nevertheless paint a picture where, in the ideal case, the case of a person who follows

their advice, our faculty of speech should become almost epiphenomenal.[56] According to these two authors, conscious speech is a faculty that should be merely used to report our mental states to others, much like the steam whistle serves to report on the locomotive's functioning. But we should not use our conscious speech to think through complex decisions. These decisions should be relegated to our unconscious thinking.

I don't think that it is possible to offer a knock-down argument against epiphenomenalism in general or against the modified form of epiphenomenalism that Wilson and Dijksterhuis propose. I think, however, that there are a number of considerations which suggest that the picture that Wilson and Dijksterhuis propose is incorrect. I will argue that our capacity to speak consciously should play a central role in helping us to make our minds (and, because of this, endorsed self-knowledge is ethically significant). To argue for this claim I will start by discussing some of the problems in Wilson's and Dijksterhuis' proposals and will then provide a positive argument defending the view that our conscious speech and our capacity to know our mental states, *qua* speakers, play a central role in our lives.

Wilson and Dijksterhuis have similar explanations for why unconscious decisions are better than conscious ones.[57] According to Wilson, when participants try to reflect on the reasons that justify their states of mind, they are often lead astray by reasons that just happen to come to mind, even if these reasons do not correspond to the person's true reasons.[58] Wilson also suggests that our conscious thought is notoriously bad at detecting correlations between different variables[59]. Both authors suggest that although conscious thought is rule-based and very precise, it cannot process large amounts of information at

56. I say "almost" because Dijksterhuis is willing to grant that at least in some cases, as when decisions are simple, we should rely on this capacity.

57. For a very different explanation of why unconscious thinking is superior to conscious thinking, see Girenzer 2007. It is beyond the limits of this chapter to explore it.

58. Timothy D. Wilson 2002, pp. 169–70.

59. Timothy D. Wilson 2002, pp. 26–7, 62.

any single time.[60] As a consequence, conscious choosers can only take into account a subset of the information required to make a proper assessment.[61] This leads Dijksterhuis to argue that, when faced with complex choices that involve assessing different attributes, we fail to weight the different attributes according to the preferences that we should consider.[62] In sum, Wilson and Dijksterhuis point to a series of factors that interfere with our capacity to think consciously. Their interpretation of the data suggests that to avoid these interferences we should make these decisions unconsciously.

My first objection to their view is that the fact that several factors impair our capacity to think consciously does not entitle them to conclude that unconscious reasoning is better than conscious reasoning. What these authors have established is that a decision based on certain bad reasons is worse than a decision based on our unconscious thinking. But this is compatible with holding that a decision based on good reasons is better than a decision based on our gut feeling. It is plausible to think (and I will argue for this shortly) that when our conscious thinking is based on good reasons, such thinking is more rational and, as a consequence, better than unconscious thinking. As a consequence, conscious thinking based on good reasons allows us to live our lives better.

Wilson and Dijksterhuis are perhaps unaware that they are, at least at times, tacitly committed to the alternate view that I just put forth. Several of the experiments that support the superiority of unconscious thinking compare the participant's choices with a "superior" choice. But the superiority of this choice is determined by conscious thought, not by unconscious thinking. Dijksterhuis, for instance, compares the car choices of participants against the car that, according to him, is the *best* given the options. But Dijksterhuis determines that this is the best car by carefully weighing the evidence, by reasoning consciously not through unconscious thinking. The fact that these authors use their conscious reasoning to

60. Dijksterhuis, Bos, et al. 2006, pp. 1005–6; Bos and Dijksterhuis 2012, p. 182.

61. Timothy D. Wilson 2002, pp. 23–9; Dijksterhuis, Bos, et al. 2006, p. 1005.

62. Dijksterhuis, Bos, et al. 2006, p. 1005; Bos and Dijksterhuis 2012, p. 182.

determine the "superior choice" shows that they are committed to the reliability of conscious thinking and to the view that when conscious thinking is based on good reasons, it is the alternative that one should choose.

Interestingly, Wilson provides the reader with tips to improve the quality of her hypothetical investigations about how she will react to situations in the future. He mentions that if we vividly imagine the potential circumstances related to what is being asked, if we carefully envisage the circumstances related to what we are meant to report, our capacity to come up with more appropriate answers will improve.[63] This tip can be taken as the starting point for an alternative proposal to the one that he explicitly favors, a proposal whereby Wilson would not be asking us to engage in unconscious thinking but would be suggesting strategies to improve our conscious self-examination and, thereby, our conscious reasoning about our mental states.

In speaking about the thought experiment inspired by Wilson's and Dijksterhuis' research I have often used the subjunctive mood. This has been my way of indicating that their results are not firmly established. An increasing number of published studies have failed to replicate the effect that Dijksterhuis allegedly established, and some recent review articles on the topic find that there is not sufficient evidence supporting some of the main theses that are meant to undergird our thought experiment.[64]

63. Timothy D. Wilson 2002, pp. 173–4. When he is sharing this "tip," Wilson suggests that this strategy helps to identify the "feelings generated by the adaptive unconscious" (Timothy D. Wilson 2002, p. 174). If one attends to the strategy that he discusses, however, one can see that it applies not only to feelings generated by the adaptive unconscious but to many other mental states as well as to the reasons which justify them.

64. See, for instance, Bonke et al. 2014; Vadillo, Kostopoulou, and D. R. Shanks 2015; Nieuwenstein et al. 2015.

IV.6.2 Empirical Research and the Ethical Significance of Endorsed Self-Knowledge

Wilson's and Dijksterhuis' results appear to suggest a certain version of epiphenomenalism according to which our conscious speech is a faculty that should be merely used to report our mental states to others, much like the steam whistle serves to report on the locomotive's passing by. According to them, we should not use this faculty to think through complex decisions. It is better to relegate these decisions to our unconscious thinking. This, however, is a view that conflicts with a whole lot of the empirical research which I discussed in chapter II. I appealed to this research earlier to show that the development of virtue required self-knowledge. In chapter II I appealed to this research to articulate why it was important to be able to self-ascribe your mental states in speech. I did not distinguish there between different ways in which you could self-ascribe these mental states. My argument ignored the distinction between first- and third-personal self-knowledge and between endorsed self-knowledge and merely-expressive self-knowledge. In what follows I will revisit some of the empirical results discussed earlier and show how they support the view that endorsed first-personal self-knowledge is of ethical significance.

Proper Reasons Are Conscious

A number of experiments have shown that performance on logical reasoning tests is heavily dependent on being able to perform these tests when one is fully attending to them, that is to say, when these tests are occurrent to consciousness.[65] Logical reasoning deteriorates sharply when one is preoccupied with other thoughts and improves when one can focus one's whole conscious attention on such reasoning.[66] As I argued earlier (II.3) this supports the

65. Neys 2006; DeWall, Baumeister, and Masicampo 2008; Baumeister and Bargh 2014, pp. 42, 45, 46.

66. Baumeister, Vohs, and Masicampo 2014, p. 21.

idea that when states of mind are occurrent they are likely to hook up with other mental states more rationally.[67]

Solving these tests will usually require intermediate steps which will involve the creation and transformation of a number of mental states: beliefs about what one has achieved, intentions about how to proceed further, desires associated with these intentions and, in some cases, emotions elicited by the engagement with the test. Every milestone achieved in the process can be seen as a moment where mental states are created or transformed. I argued before that when the participants are distracted, many of these mental states (and many of these intermediate steps) will be unconscious. Conversely, when participants are able to pay full attention to the test, these intermediate steps and the mental states associated with them will tend to be conscious. All of this was discussed in chapter II. What I want to point now is that these occurrent mental states that help the person come up with a solution will be known with first-personal self-knowledge. In fact, they will be known with endorsed self-knowledge. These mental states arise through the person's self-conscious efforts to solve the test. As a consequence, their content is something not merely on which the person can report, but something that she will endorse *qua* speaker, as an intermediate step in the solution of the test.

Empirical research also suggests that making long-term plans requires that we are able to consciously entertain these plans.[68] Once again, thinking through a plan will entail an agential perspective on the part of the person. The different considerations that come up in the context of designing this plan are ones to which the person will not be able to report as a witness; they are considerations which the person will endorse. Because these considerations are ones on which the person is relying to develop the plan, they are such that the person actually needs to endorse them.

67. Baumeister and Bargh 2014; Baumeister, Masicampo, and Vohs 2011; Gawronski and Strack 2004.

68. Baumeister and Bargh 2014, p. 43. See also: Baumeister, Masicampo, and Vohs 2011, p. 336.

This provides a first empirical confirmation that when you make up your mind both the process and the mental states elicited by such process will tend to be more rational when you are able to report on them with endorsed first-personal self-knowledge. And if we grant that being more rational is valuable for ethical development, that it is part of a virtuous life to have mental states that are rational and respond not merely to associative patterns but to properly justified warrants, then it follows that endorsed self-knowledge is valuable for ethical development.

Living a Unified Life According to Long-Term Ethical Principles

I also mentioned that unconscious thinking is primarily aimed at producing adaptive responses in the immediate present.[69] As such, unconscious thinking is not capable of considering multiple perspectives or long-term horizons; it is, rather, oriented by what Freud characterized as the pleasure principle.[70] As a consequence it is unlikely that unconscious thinking will be able to come up with responses that correspond to the ethical aspirations of the person, at least to those which have to do with long-term considerations about how she is to live her life as a whole. The empirical results from this area of research, then, suggest that being able to take a broad perspective where one considers one's life as a whole requires that we are able to think consciously about it. Once again, this capacity to think consciously is not merely a capacity to self-ascribe one's mental states detachedly. Because this is a perspective that presupposes that we are endorsing what we are thinking, it entails that it requires that we have a particular type of self-knowledge: endorsed self-knowledge.

Research has also suggested that when mental states conflict (say when you have two desires which clash), being able to occurrently hold these conflicting states of mind is impor-

69. Baumeister and Bargh 2014, p. 47.

70. Freud 1958f; Timothy D. Wilson 2002, p. 38.

tant for mediating and resolving the conflict.[71] These two desires ought to be held together in order to come to a decision. And this entails that they have to be known with endorsed first-personal self-knowledge.

Both of these results suggest that when Chloe allows her unconscious thinking to decide whether to marry Mark or not, she will not be evaluating her decision in light of her long-term ethical principles. Instead, she will be responding to needs that are more immediate and which obey to baser inclinations. Furthermore when her unconscious thinking needs to make a choice between different competing inclinations, either it will not be able to mediate between them (leading Chloe to sabotage her own actions) or it will merely arbitrarily pick one or the other without giving proper consideration to the complexity of the conflict.

The Unconscious Does Not Understand Our Language

Nearly all psychologists nowadays agree that occurrent thought is required for verbal communication. I mentioned that unconscious thinking is blind to negation, it is not able to detect complex ideas formulated in language, and it is not able to properly parse the syntax and semantics of sentences. Properly parsing and understanding fully formed sentences requires occurrent thought.[72]

All of this undermines the idea that Mark and Chloe can have a conversation or discussion, even the weird conversation about marriage that we entertained when we explored the thought experiment inspired by Wilson and Dijksterhuis. If Chloe decides to think unconsciously about Mark's proposal (i.e. "You will manage our bank account and all the money I make will go into it") she will likely fail to properly track what he is saying. Unconscious thinking gets primed by individual words. In this sentence the combination of the words

71. Baumeister and Bargh 2014, p. 44.

72. See, for instance, Baumeister, Vohs, and Masicampo 2014, p. 21; Baumeister and Bargh 2014, p. 40; Gendler 2008a, p. 649; Timothy D. Wilson 2002, p. 65.

"bank account" and "money," together with "managing" and "making" is likely to prime associations that will present Mark as really concerned with money, with "making it" and "managing" it. If so, it will tend to lead Chloe's unconscious thinking to associate Mark's remark with her prior belief that he is too thrifty; his remark would be processed by her unconscious thinking in exactly the wrong way.

Thus, it is only when a person has endorsed self-knowledge that she will be able to discuss her own views with others (and the mental states which reflect these views). If a person lacks endorsed self-knowledge, then she will be at a distance from any mental state that she self-ascribes. If you are a mere witness of the mental states of someone else, learning what speaks for or against that mental state will be incapable of changing this person's mental state. And this is so even if this someone else happens to be you. To suppose that the criticisms to the mental states that you know with third-personal self-knowledge will have an effect on these mental states is to suppose that you are able to transform these mental states without the mediation of you, *qua* speaker, that you are able to transform these mental states unconsciously. But if your unconscious thinking has a very bad ear to linguistic utterances, as the research suggests, then it means that if these mental states are transformed, they are not transformed as a consequence of the recognition of the warrants (or lack thereof) that justify holding such mental states. If these mental states change, it is as a consequence of rather blind and associative processes.

So if one believes, like Darcia Narvaez does, that "the moral life involves co-authoring the future with others through dialogue and feedback on imagined alternatives,"[73] then it follows that endorsed self-knowledge will be necessary for ethical development. Talking about our desires, beliefs and emotions allows us to get different perspectives from interlocutors different from us, perspectives that can serve as a corrective to our blind spots and biased conceptions. Only with endorsed self-knowledge will you be able to transform, in a non-

73. Narvaez and Mrkva 2014, p. 26.

accidental way, your mental states through dialogue with and feedback from others.[74]

Speaking (and Writing) About Our Mental States

I also mentioned that there is a robust body of research showing that a person's mental health improves when she speaks (or writes) about thoughts and emotions related to traumatic experiences that she has experienced.[75] Although the experiments have not explicitly controlled whether the participants are speaking about them as witness or as a committed agent, it is plausible to suppose that most of the people who speak or write about their thoughts and emotions do so with endorsed self-knowledge. The research has established that speaking about these mental states is not beneficial when participants merely rehearse or relive the experience; it is beneficial when they are able to analyze and investigate it. One of the central results from this research is that speaking (or writing) about these issues contributes to mental health when the person is able to construct a new narrative that integrates this experience within her life as a whole. But this creative process, whereby one's reflection on one's emotions leads to a transformation and reframing of them is, precisely, the kind of dialectic process between detection and creation that characterizes endorsed self-knowledge. It is plausible to suppose that being able to report these mental states with endorsed self-knowledge is what makes it possible to transform many of these mental states from inchoate intuitions that guide the person, either unconsciously or unreflectively, into properly rational responses, embedded in properly reason-giving relationships.

74. I say "in a non-accidental way" to contemplate the possibility that lose associative processes, primed by the words that occur within these conversations, might happen to transform the mental states in such a way that, by a happy accident, they end up reflecting more rational responses to the world.

75. Pennebaker and Chung 2014, p. 418.

IV.6.3 Third-Personal Self-Knowledge Undermines the Authority of Your Capacity to Reason Consciously

Wilson and Dijksterhuis offer a distinct theoretical possibility that helps us see that defenders of agency models of self-knowledge rely on an unargued premise, namely, that our rationality is to be identified with our capacity to speak. This identification, as I have just argued, is actually true about human beings. But the identification depends on empirical arguments about the way in which speech works in creatures like us. Human beings use (and *should* use) their speech not merely to report on their mental states but to make judgments on the world, judgments which, in turn, contribute to the constitution of their mental states. And although it is natural to think that, in any creature endowed with the capacity to speak, speech would fulfill this same function, it is theoretically conceivable that there could be creatures for which this faculty could be a hindrance, one that they should avoid when they had to make up their minds, one that they should use merely to offer witness reports on their mental states.

McGeer suggests that "third-personal self-knowledge undermines the authority that you have to make up your mind."[76] Everything I have said so far should help us see that the right way to put her point is as follows: "third-personal self-knowledge undermines the authority that you have, *qua* speaker, to make up your mind." Putting it in this way makes clear the fact that your mind is often made up without your conscious reasoning, without the intervention of *"the speaker in you."* This, in turn, makes perspicuous the fact that your mind can be made up in different ways. In other words, it makes perspicuous the theoretical possibility inspired by Wilson and Dijksterhuis that I discussed above.

Once this is on the table, it becomes clear that defending the ethical significance of first-personal self-knowledge involves defending the ethical significance of this particular way

76. McGeer 2007, 81. This is not a literal citation.

to make up our mind, of making it up through the mediation of *the speaker in you*. The empirical arguments I just provided were meant to provide such a defense by showing that the capacity to make up your mind *qua* speaker is the capacity required to: 1) make up your mind in a properly rational way; 2) take a long-term perspective where you mental states are a response to the question "how should I live?"; and 3) allow you to discuss and think about the response to "how should I live?" in the company of others. Aspiring to achieve only third-personal self-knowledge, relying on decisions based only on third-personal self-knowledge, undermines your capacity to reason consciously, a capacity that, because of the three claims I just made, is essential to live a flourishing life.

A caveat is in place. Sometimes it may be ethically valuable for a person to undermine her authority to reason consciously. For some people it might be better not to have a well developed capacity for first-personal self-knowledge. As I said in chapter II, ignorant virtue is better than knowing viciousness and better than misguided conscience. If your aspirations are misguided or if your capacity to reason consciously malfunctions and leads you to make up your mind in the wrong way, endorsed self-knowledge might actually lead to develop vice and not virtue. All of this suggests a further way in which we need to pare down the view defended by McGeer. Your authority *qua* speaker should not, in all circumstances, be used to make up your mind.

But even when your ethical aspirations are on the right track and even if your capacity to consciously reason works properly, it is sometimes ethically valuable to surrender your capacity to reason consciously. There are occasions where wiser people can see much better than we do what it is that we should believe, feel or do. In these cases, following the wise person's advice will lead to behavior that accords more with virtue.

Within most accounts of virtue, however, behaving in accordance with virtue is insufficient to behave virtuously. Behaving as a wise person tells us is often insufficient to align our whole mental life in the appropriate way. Virtue is often said to involve, not merely the

execution of a particular deed, but an emotional and cognitive alignment with such deed. And this requires not merely that we are able to follow instructions from someone wiser, but that we can see and respond like the wise person does. This is, of course, not the only problem with acting merely in accordance with the advice of a wise person. Autonomy involves our capacity to act from our own understanding of how we should think, desire, act and feel. Even though following the advice of people wiser than ourselves is a central part of our ethical development, the ultimate aspiration of the ethical pilgrim is to be able to become wise enough that she can rely on her own assessment of the merits of the situation. Anything else than this puts into question the idea that a flourishing life is lived by autonomous and self-determining creatures. This provides us with a further way in which McGeer's claim needs to be pared down. There are occasions where we ought to relinquish our capacity to make up our minds by reasoning consciously.[77]

IV.7 Overvaluing Third-Personal Self-Knowledge Undermines Self-Knowledge

I argued in prior sections of this chapter that first-personal self-knowledge was both pervasive and ineliminable. I'd like to conclude this section by showing that if one believes that third-

77. It is important to note a potential ambiguity in the statement "make up your mind through your capacity to reason consciously." On the one hand, one could interpret this in such a way that the decision to follow the advice of a wiser person, even against your own assessment of the facts of the matter, does not entail surrendering your capacity to reason consciously. According to this interpretation, it is your recognition of this person's wisdom that leads you to reach the all-things-considered judgment: "Follow her advice, regardless of your independent assessment of the merits of the situation." On the other hand, there is an alternative way to interpret this idea. According to this second interpretation, the all-things considered judgment "follow the wiser person's advice regardless of your independent assessment of the merits of the situation" *is* a relinquishment of your capacity to make up your mind consciously. This is so because you might conceive of this capacity as the capacity to make up your mind based on a recognition of the merits of the situation, not on the authority of what a third person might say about this.
McGeer's criticisms of contemplative conceptions of self-knowledge are of a piece with her exhortations to be autonomous self-determining individuals. If these two sets of ideas are to hang together, we need to understand what it means to "make up one's mind" in the second way, a way which excludes our relying on the authority of other people to determine how we ought to live.

personal self-knowledge is sufficient for ethical development one ends up undermining the idea that self-knowledge is important for ethical development.

We see this by returning to our thought experiment and noticing that Chloe's self-knowledge actually did not play any essential role in her interactions with Mark. Chloe was merely repeating what the robotic voice of the Inner Self-Detector was saying. For all intents and purposes, Mark could have just relied on what the Inner Self-Detector said; he did not require Chloe to repeat it. Whether she knew this information or not was irrelevant in their interactions. Thus, it was not important for Chloe to know her mental states.

One might object that even if Chloe's self-knowledge was irrelevant in her conversations with Mark, there are other occasions where it would be important for her to come to know her mental states. Think of the case where she needs to grab some food for lunch. Because she is on top of the empirical literature she knows that she will be happier if her explicit and implicit motivations align.[78] Thus before making up her mind consciously about what and where to eat she asks the Inner Self-Detector: "What kind of food do I want to eat today?" The robotic voice responds: "You want Mexican." This response will determine her choice of restaurant. And one might argue that it is only because she came to know about this fact that it was possible for her to determine where to go.

In response I want to point out that, in this case, the Inner Self-Detector's response does not merely provide Chloe with third-personal self-knowledge. When the voice responds: "you want Mexican" Chloe comes to know something about herself. But to the extent that this piece of self-knowledge helps her decide where to go, it becomes a piece of self-knowledge which she endorses. If Chloe was a mere witness of her desire to eat Mexican this desire would not be able to inform a practical deliberation that concludes in an action to drive to *La Casa del Pueblo*. It would be, at most, a hypothetical exploration about where the person who happens to be her would drive. And this just means that claiming that self-knowledge is

78. I talked about this in II.4.3

important to ethical development entails granting that this self-knowledge cannot be merely third-personal.

IV.8 Importance of Third-Personal Self-Knowledge

Even though I have suggested that third personal self-knowledge is therapeutically inadequate, it is important to recognize (and emphasize) that it is not inert. In fact, I want to conclude this chapter by briefly mentioning a few ways in which third-personal self-knowledge is actually valuable for ethical development. Of course, given what I have said before, third-personal self-knowledge will be valuable, either as a stepping stone to acquire first-personal self-knowledge, or because its objects cannot be known with first-personal self-knowledge.

IV.8.1 Therapy

In the *Introductory Lectures* Freud himself acknowledged that while informing the patient about the sense of her symptoms does not have the result of removing them, it does get the analysis in motion.[79] In "Wild Psychoanalysis," while criticizing the procedure of 'wild psychoanalysts' who rushed to provide analysands with an account of their underlying symptoms, he conceded that "a clumsy procedure like this, even if at first it produced an exacerbation of the patient's condition, led to a recovery in the end. Not always, but still often."[80]

Third-personal self-knowledge, then, often has the result of putting in motion certain psychic processes that will bring about first-personal self-knowledge and ethical self-development. And even if the ultimate aim of the ethical pilgrim's aspirations to know her mental state is to know them with first-personal self-knowledge, third-personal self-knowledge is sometimes

79. Freud 1963, p. 281.

80. Freud 1995c, p. 227.

a necessary step to acquire it. There are mental states that the ethical pilgrim needs to know but which are difficult to acknowledge because they are too anxiety provoking. In these cases, third-personal self-knowledge affords a certain emotional distance that allows the person to start acknowledging that she holds these. Thus, when mental states are very charged emotionally, third-personal self-knowledge is often instrumental to the ethical pilgrim's gradual development of first-personal self-knowledge. It allows her to entertain the possibility that she holds them, a possibility that will then pave the way for her acknowledging these mental states in a first-personal way.

IV.8.2 Scientific Self-Knowledge

There are a number of things about ourselves that cannot be known first-personally but which are important for ethical development. Timothy Wilson suggests that it is valuable for us to know about ourselves by becoming acquainted with the scientific literature: "Many people learn about their bodies by reading about medical research, such as studies on the dangers of tobacco, saturated fat, and ultraviolet radiation. Given that we have no direct, privileged access to how our pulmonary or cardiovascular systems work, we are at the mercy of such out-side sources of information to inform us about how things like smoking tobacco influence our health. I suggest that the same is true in the psychological realm."[81] As Wilson argues, a significant number of things that are known this way simply cannot be known first-personally (even if some of these things have to do with our own minds).

IV.8.3 Aliefs, Phobias and Unconscious States of Mind

I have hinted at the fact that part of the importance of endorsed self-knowledge is that this form of self-knowledge allows us to transform our mental states in a rational way (and not accidentally so). It is worth mentioning, however, that even if this is true, it is still the case

81. Timothy D. Wilson 2002, p. 183.

that there are a number of mental states that need not or sometimes cannot be transformed in this way.

For instance, there are certain interventions where Behavioral Therapy has been proven to be quite successful as a strategy to change a person's mental states. Phobias, in particular, are often treated successfully by merely exposing the patient to the feared object. One might argue that the person's fearful responses are not transformed by the mediation of the person's conscious reflection and, as such, that this kind of self-knowledge does not require the mediation of endorsed first-personal self-knowledge.

There are also a number of mental states that appear to correspond to what Tamar Gendler has called aliefs, namely, mental states that are "associative, automatic, and arational."[82] Aliefs are mental states that are not responsive to reasons but only to brute associations. Because of this, we should not expect rational reflection to be efficacious to transform them. If the ethical pilgrim wants to change an alief, then she will most likely need to change it by establishing new associations. It is hard to see how she could acquire endorsed self-knowledge of this mental state nor how this self-knowledge could be instrumental in changing it.

Finally, there is plenty of clinical evidence that suggests that an important part of the process of improvement within the therapeutic setting has to do with unconscious processes where the person's consciousness does not interfere. In their paper on therapeutic action, Glen Gabbard and Drew Westen claim: "Although many of the avenues to change described by contemporary theorists involve explicit interventions, conscious mastery of the implicit and repetitive modes of relatedness is often accompanied by changes in non conscious affective and interactive connections described by Lyons-Ruth et al. (1998) as implicit relational knowing."[83] Not all of the transformation in our mental states requires our conscious inter-

82. Gendler 2008a, p. 641.

83. Gabbard and Westen 2003, p. 825.

vention. To that extent, it is not true that all of the transformations in our mental states require us to have endorsed self-knowledge.

IV.9 Conclusion

In the first sections of this chapter I defined and clarified the main types of self-knowledge that will occupy us in what follows: first-personal, third-personal, endorsed and expressed. While laying the conceptual groundwork on which the rest of the dissertation depends I articulated explicitly the relationship between first/third-personal forms of examination discussed in the previous chapter and first/third forms of knowledge that were discussed in this one. I argued that there was a noteworthy asymmetry: while first-personal self-*examination* only leads to first-personal self-*knowledge*, third-personal self-*examination* might (and sometimes should) lead to third-personal self-*knowledge*.

I then spent the rest of the chapter elucidating why the ethical pilgrim should aspire, not merely to acquire information about her mental states, but to know her mental states in such a way that her own reflections on them and their merits contributed to constituting these mental states. I showed that endorsed self-knowledge is ineliminable and pervasive. Furthermore I articulated why possessing endorsed self-knowledge contributes to develop virtue. The mental states that constitute your mind will tend to be more rational when you are able to report on them with endorsed self-knowledge; lacking endorsed self-knowledge of your mental states will also tend to make them less amenable to be responsive to your long-term ethical principles and to respond, rather, to needs or desires that are more immediate and often obey to baser inclinations. Finally, being able to report these mental states with endorsed self-knowledge makes it possible to transform many of them from inchoate intuitions that guide the person, either unconsciously or unreflectively, into fully-fledged responses that are embedded in properly reason-giving relationships. In sum, endorsed self-

knowledge allows the person to unify capacities that are central to a person's flourishing: the capacity to speak, to reason and to act.

I'd like to close this chapter by mentioning how my account bears on traditional theories of self-knowledge within epistemology. The writings of philosophers of mind about self-knowledge have mainly focused on offering an account of the allegedly privileged, immediate and authoritative knowledge that one has about one's intentional attitudes. My aim in the dissertation is not to explain this knowledge but to understand how self-examination and self-knowledge contribute to the development of virtue. It just so happens that there are insights in the debates by traditional epistemologists that are applicable to the questions which I am pursuing here.

But even if my project is different from the project that epistemologists pursue under the heading "self-knowledge," what I say here is not completely orthogonal to what they write. The results of my investigation determine some constraints that any adequate theory of self-knowledge should respect. To begin with any appropriate theory of self-knowledge should be able to account for the distinctions between first- and third-personal self-knowledge, on the one hand, and for the distinction between endorsed self-knowledge and merely-expressive self-knowledge, on the other. It also follows from my account that any adequate theory of self-knowledge should make room for the fact that there is a dialectical relationship between detecting, confirming and transforming our mental states. That is to say, any account of self-knowledge has to account for the fact that judgment-sensitive mental states are not inert items that lie in our minds to be passively contemplated but active elements that are responsive to the world, updated and transformed by our reflections, judgments and any new information that we come to have about it. Thus, any theory of first-person authority ought to be able to account for the fact that the judgment that I have about my mental state M is not independent of my own judgment that it is warranted to hold M.

V

Merely-Expressive First-Personal Self-Knowledge

Although most virtue ethicists hold that self-examination and self-knowledge of our mental states play an important role in the development of virtue, very little attention has been paid to understanding this role. Moral philosophers rarely mention that the aspiration to know our mental states, if it is guided by ethical considerations, is not merely an aspiration to acquire information about these mental states; it is an aspiration to relate to these states in a manner that makes it possible to engage with and transform them in a distinctive way.

That the distinction between different ways to know your mental states is significant for ethical development has been explicitly acknowledged by most schools of psychotherapy. It took Sigmund Freud more than a decade, and a number of therapeutic failures, to recognize that not all forms of self-knowledge are equivalent therapeutically. According to him, it is not difficult for an experienced doctor to know the mental impulses that have remained unconscious in a particular patient, impulses which, according to him, she should come to know if her condition is to improve. Freud emphasizes that if the patient comes to know about the sense of her symptoms through the testimony of her analyst, this knowledge will

often be inadequate to transform her.

> Knowledge is not always the same as knowledge: there are different sorts of knowledge, which are far from equivalent psychologically. (. . .) The doctor's knowledge is not the same as the patient's and cannot produce the same effects. If the doctor transfers his knowledge to the patient as a piece of information, it has no result.[1]

In the early stages of the dissertation I thought that I could cash out this Freudian insight by distinguishing between two ways to know one's states of mind: first-personally and third-personally. As I worked on the topic, however, I came to realize that the situation was more complex. To properly understand the role of self-knowledge in ethical development one needs to draw an additional distinction between two types of first-personal self-knowledge.

In the previous chapter I examined endorsed self-knowledge. I argued that the possession of endorsed self-knowledge puts you in a position to make up your mind in a properly rational way, to live your life within a long-term perspective that understands it as the answer to the overarching question "how should I live?" and to reflect on this response in the company of others. This form of first-personal self-knowledge is a reflection of the subject's unity as speaker, reasoner and doer. In the current chapter I highlight the place that expressed self-knowledge plays in ethical development by examining the complementary ways in which expressed and endorsed self-knowledge contribute to therapeutic improvement. To do so I will examine the contrasting accounts of self-knowledge offered by Richard Moran and Jonathan Lear, accounts that offer a conception of self-knowledge that correspond, respectively, with what I have called endorsed self-knowledge and merely-expressive self-knowledge.

Moran takes his account of self-knowledge to reflect the self-knowledge that, according to Freud, patients where meant to achieve in therapy. Lear criticizes Moran both for his account

1. Freud 1963, p. 281.

of first-personal self-knowledge and for his characterization of psychoanalysis. Lear helps us to see that the mental states that the analysand is meant to come to know first personally are often recalcitrant to reason. The kind of self-knowledge that Lear puts forth is knowledge of patterns of thought that interfere with the person's ability to live well. Moran fails to see that these recalcitrant mental states, which we cannot know with endorsed self-knowledge, are at the heart of Freud's therapeutic practice. This failure leads him to mischaracterize what takes place within therapy. But while Moran fails to see the instrumental role that merely-expressive self-knowledge plays in therapy, Lear appears to be blind to the normative role of endorsed self-knowledge. Possessing endorsed self-knowledge is a sign of having a healthy mind, and an ultimate aim of the ethical pilgrim. Lear seems not to recognize that the self-knowledge that he discusses is defective and, as such, cannot be the ultimate aim of the ethical pilgrim's aspiration to know her mental states.

I will conclude the chapter by contending that Lear's and Moran's accounts should not be seen as competing but as complementing each other. Each of them help us understand different aspects of the self-transformation that takes place in psychoanalysis. Endorsed self-knowledge sets up the ideal that should be aimed at in therapy, while merely-expressive self-knowledge provides an important mean that helps us attain that ideal.

Because Moran's and Lear's discussions focus on a traditional form of Freudian psychoanalysis, in this chapter I will do the same. I will be also taking it for granted that psychotherapies have valuable things to teach to philosophers interested in self-transformation. This is so, not just because these practices aim to help individuals transform themselves but because the desired transformation is an ethical one. This last idea might seem controversial because of the widespread assumption that contemporary forms of therapy are morally neutral and do not deal with ethics. A number of philosophers, however, have shown that, despite psychotherapist's own denials, their practices are, at their core, shaped by ethical

aims.[2]

V.1 Transparency and Endorsed Self-Knowledge

V.1.1 Transparency and Avowal

In *Authority and Estrangement* Moran writes:

> [There is a] crucial therapeutic difference between the merely "intellectual" acceptance of an interpretation, which will itself normally be seen as a form of resistance, and the process of working-through that leads to a fully internalized acknowledgment of some attitude which makes a felt difference to the rest of the analysand's mental life. This goal of treatment, however, requires that the attitude in question be knowable by the person, not through a process of theoretical self-interpretation but by avowal of how one thinks and feels. That is, what is to be restored to the person is not just knowledge of the facts about oneself, but self-knowledge that obeys the condition of transparency.[3]

In the next section I will explain why, according to Moran, "restoring to the person self-knowledge that obeys the condition of transparency" is central to mental health. Before doing so, however, we will need to understand what Moran means by "transparency," a locution that, within the context of the dissertation, can be quite misleading.

There are different ways in which a person can know that she holds a certain mental state. A person can be a careful observer of her behavior, inferring from this behavior

2. See, for instance, Sherman 1995, Martin 2006, Lear 2006 and Lacewing 2013. The core argument defending this view is that the notion of therapeutic improvement only makes sense within a larger notion of flourishing or virtue.

3. Richard Moran 2001, pp. 89–90.

that this mental state can be attributed to her. In the case of a mental state like a belief, however, a person can also find out what she believes by looking out onto the world. When you ask a person: "Do you think that it will rain?," she can respond to this question by attending to the same outward phenomena to which she would attend if she were to answer the question "Will it rain?" Although both of the above questions are internally related, the question "Do you believe that P?" asks for something different than the question "In light of the evidence, is P true?" The first is a question about the person, about her own mental states. The second question is a question about the world, answered by the person by looking outwards, at the facts with which P is concerned, facts that are meant to provide evidence that warrants P.

The concept of transparency, as Moran uses it, does not merely apply to beliefs. It is meant to apply to any mental state that can be formed by examining facts about the world, facts that might make no explicit reference to us. Thus, transparency is a concept that has application in the genus of judgment-sensitive mental states. For any judgement-sensitive mental state it is true that a person can say whether she has it by examining facts about the world. A person can often respond the question: "Do you have an intention/emotion/desire?" by attending to the same outward phenomena to which she would attend if she was answering the question "Is this intention/emotion/desire merited?" (or "Is this intention/emotion/desire one that I ought to hold?")[4] The person's judgment "This mental state is merited" is meant to be non-accidentally connected with the fact that this person's has such mental state.[5]

4. This, of course, is not to say that this is the typical way in which we respond to this question. My claim here is only that this is a potential way to respond to it.

5. The formulation of transparency that I have just offered might be thought to be different from Moran's own exposition. In his exposition he starts with the now a famous example by Gareth Evans: if you ask a person: "Do you think that there is going to be a third world war?," the person will often respond by attending to the same outward phenomena to which she would attend if she was answering the question "Will there be a third world war?" (G. Evans 1982, p. 225, quoted in Richard Moran 2001, p. 61). Moran characterizes transparency as the relation that holds between two questions: "Do you believe P?" and "Is P true?", not between the questions "Do you hold P" and "is P warranted?" But although Moran characterizes

The locution *"transparency"* is meant to remind us that when the person wants to find out what her mental state is in this way, her gaze does not aim at, or focus on, her mind. Rather, it looks through her mind, at the facts to which the mental state refers. In looking outward this person makes a judgment about the merits of holding a particular mental state. If she finds it to be merited, (and if her mental processes are working properly) this judgment will be non-accidentally connected with her possession of this mental state.

Transparency is a term of art for Moran. In the context of this project it is a term of art that can be quite confusing. When discussing self-examination and self-knowledge, the adjective "transparent" is typically used to characterize mental states that are not opaque to the person, mental states whose content and nature are unproblematically available to her. But transparency, *à la* Moran, is a concept that contrasts, not so much with opacity, as with irrationality. To see this, think of the case of an irrational phobia of spiders which the person recognizes as unwarranted. This person's phobia might be easily accessible to her. But because this person's fear of spiders is not responsive to her assessments of the merits of fearing spiders, her fear will not be transparent in Moran's sense.[6]

V.1.2 Potentially or Effectively Judgement-Sensitive

It might be worth reminding the reader that the property of being "judgement-sensitive" characterizes states of mind that *ought* to be responsive to our judgments even if, as a matter of fact, they aren't actually responsive to such judgments. Saying that a mental state is

---

transparency as the relation that holds between the first two questions, he makes it clear that we should hear the latter of these first two questions in a particular way, as the question "Ought I to believe P?" In fact, he explicitly connects the question about the belief of P with the question the merits of such belief: "What the 'logical' claim of transparency requires is the deferral of the theoretical question "What do I believe?" to the deliberative question "What am I to believe?" And in the case of the attitude of belief, answering a deliberative question is a matter of determining what is true" (Richard Moran 2001, p. 63).

6. It is worth saying that the ordinary conception of transparency (where transparency is opposed to opacity) is not entirely different from Moran's transparency in two ways. 1. When a mental state is transparent, *à la* Moran, it is non-opaque. 2. When a mental state is fully transparent, *à la* Moran, attending to the passion is usually sufficient to change the passion. The question 'what do I feel?' gives way to the question 'what ought I to feel?'. And the answer to the latter determines the answer to the former.

"judgement sensitive" does not entail that such mental state has been actually formed by a judgment about its merits nor that it is, in fact, responsive to such judgment. It only implies that it is the type of state that *could* be formed and *should* be transformed by such a judgement. A recalcitrant and irrational emotion or desire that the person recognizes as such is "judgment-sensitive" because, even if it has not been formed, nor can be transformed, by the person's judgments, it is the *type* of state that, given its nature, should be responsive to these judgments. "Transparency," *à la* Moran, characterizes those judgement-sensitive mental states that, as a matter of fact, are effectively responsive to our judgments, states of mind that can be formed and will be transformed by our judgments about the merits of holding them.

Occasionally I will refer to these mental state that are effectively responsive to judgment as "transparent mental states." I will use this locution sparsely to prevent confusing between "transparency," *à la* Moran, and "transparency" in the sense of non-opaque. I will mostly refer states of mind that are effectively responsive to our judgment as "effectively judgment-sensitive" (although some times I will use the locution "judgment-responsive.") I will refer to the *property* that characterizes these states as "judgment-responsiveness" (the, perhaps most perspicuous locution, "effective judgment-sensitivity," seems too clunky to me). The taxonomy below illustrates the way in which these concepts hang together.

This tree makes perspicuous that "judgment-responsive mental states" are a a species of the genus "judgment-sensitive mental states." The structure between this genus and this species is an instance of what Anton Ford has called essential generality. It is an example where the species singles out the proper specimen of its genus.[7] This relationship is like the relationship between gold and pure gold. All pure gold is gold, but not all gold is pure. An object made of pure gold is not merely a token of the genus gold, it characterizes the proper exemplar of its genus: if you really want to understand what gold is, you should take a look

7. Ford 2011.

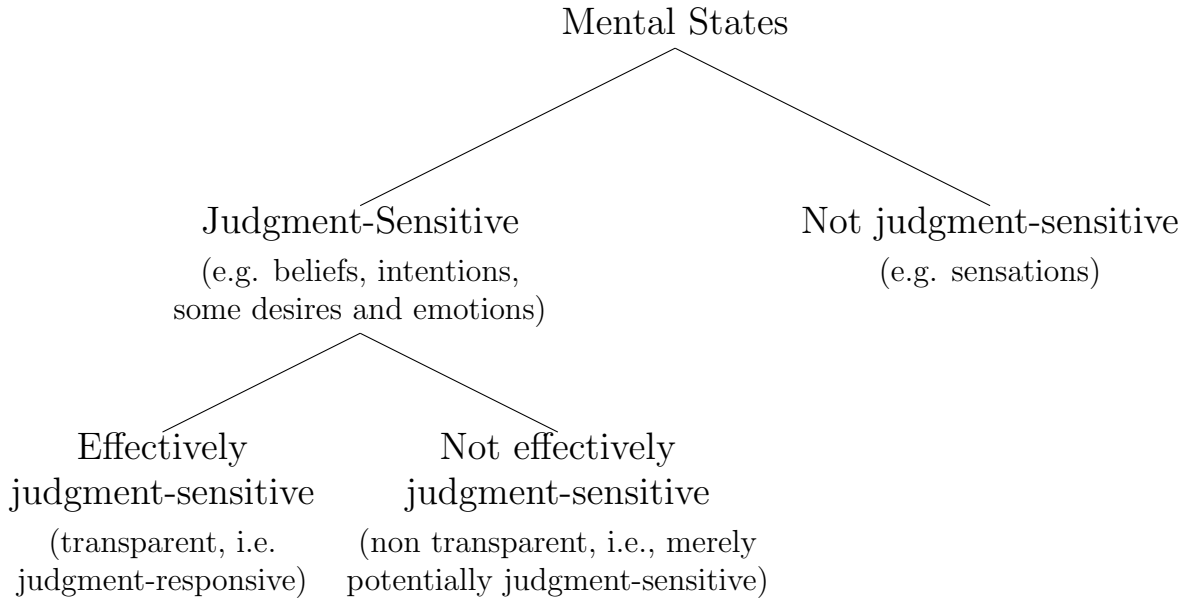

Figure V.1: A Taxonomy of Mental States

at the essential species, at pure gold. Similarly here: if you really want to understand what a judgment-sensitive mental state is, look at the essential species, i.e. at mental states that are effectively judgment-sensitive.

V.1.3 Self-Knowledge That "Obeys the Condition of Transparency"

The paradigmatic mental states that illustrate what Moran calls "self-knowledge that obeys the condition of transparency" are those where we come to have as a result of our judgments about the merits of holding them. If things are in order, when I look out and judge that, given the evidence, P is to be believed, this judgment does not only secure my knowledge that "P is to be believed," it also secures my knowledge that "I believe P." That is to say, in judging "P is to be believed" I become someone who believes P, and I know that I believe that P because I judge that P is to be believed. This account is meant to generalize to all judgment-sensitive mental states. If things are working well, when I look out and judge that, given the available evidence, it is merited to hold a mental state S, this judgment does not only secure my knowledge that I should hold S, it also secures my knowledge that "I hold

S." (If all is going well, judging that something is to be done is all it takes me to intend to do it; judging that I should be angry is all it takes for me to be angry.)

These paradigmatic examples of transparent self-knowledge might tempt some readers to think that Moran is suggesting that the psychoanalytic patient should endeavor to know her mental states by making judgments about the merits of holding them. This is a misreading and a misunderstanding. Moran is comfortable with the idea that few of our judgment-sensitive mental states come into existence as the conclusion of an explicit exercise of our capacity to judge on their merits. He is also not suggesting that we need to always be explicitly judging such merits.[8] In fact, Moran explicitly concedes that updating and revising our beliefs about our environment, not only *does not require* our explicit judgments about these merits, but *could not* require them: "In our biological as well as our cognitive lives, it must be the case that the majority of these processes take care of themselves, however 'rationalizable' they may be in their functioning. Otherwise we couldn't get going with the more reflective activities of the whole person like deliberate action or self-criticism."[9]

Moran is not proposing that the aim of psychoanalysis is to get the person to acquire all of her judgment-sensitive mental states through an explicit process of deliberation about the warrant's merits. What Moran proposes is that therapy should put a person in a position where she *can* address any of these mental states in this way, allowing evidence to inform and transform her mental states by examining (or re-examining) the warrants that merit or do not merit holding a particular mental state. What Moran is suggesting, then, is that after a successful therapy, when a person states that she fears something, she is doing more than merely giving a report on someone's fear, she is also putting herself on the line for being afraid, she is confirming that she stands behind her fear, endorsing (or re-endorsing) the fact that she considers her fear to be warranted. Thus, even if the person's report on

8. Richard Moran 2001, 116. See also p. 63.

9. Richard Moran 2001, p. 111.

her mental state is not the result of her judgment about the merits of having it, it is open to be transformed by such judgments. The property of judgment-responsiveness (i.e. of being transparent or effectively judgment-sensitive) is meant to signal that in a well functioning mind there is an internal and dynamic connection between a judgment-sensitive mental state and our capacity to reflect on it.

> "[the idea of Judgment-responsiveness] lies in the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens to be me. This is not to say that one normally arrives at one's beliefs (let alone one's fears or regrets) through some explicit process of deliberation. Rather, what is essential in all this cases is that there is logical room for such a question, about regret as much as about belief, and that the actual fear or regret one feels is *answerable* to such considerations."[10]

Moran's proposal, then, is not that mental health is attained when we know all of our states of mind by making explicit judgments about the merits of their warrants. For our states of mind to be judgment-responsive it is not necessary that these states have been formed as a result of these judgments. All that is required is that, if we ask the outward oriented question, they will effectively respond to such judgments.

What I have said should be sufficient to recognize that "self-knowledge that obeys the condition of transparency" coincides with what I defined, in the previous chapter, with "endorsed self-knowledge." As I have been saying, according to Moran, a person has this kind of self-knowledge when her self-ascription of her mental state M allows her to do more than just make a witness report of M. The person with "self-knowledge that obeys the condition of transparency" is not merely capable of ascribing M to the person who happens

10. Richard Moran 2001, p. 63.

to be her; she is able to make an endorsed report, to self-ascribe her mental state in such a way that she can put herself on the line for holding M, confirming that she stands behind it. But this just means that the person is able to express her endorsement of M (where such endorsement involves the person's recognition that such a mental state is warranted), i.e. that she has endorsed self-knowledge.

V.1.4 Endorsed Self-Knowledge and Judgement-Responsive Mental States

It might be worthwhile to make explicit the intimate connection that there is between the set of mental states that is effectively judgment-sensitive and the set of mental states that are known with endorsed self-knowledge. It might be natural to think (and this appears to be what Moran holds) that these two set of mental states are identical, that if one has endorsed self-knowledge of a mental state M, then M is effectively judgment-sensitive (and, conversely, that if M is effectively judgment-sensitive , then it is known with endorsed self-knowledge). As I will now argue, even though there is an intimate relationship between these two sets of mental states, their relationship is not one of identity.

Let me start by mentioning the exceptional cases where one has endorsed self-knowledge of a mental state M but M is not necessarily effectively judgment-sensitive. A person has endorsed self-knowledge of a mental state M when he has the capacity to self-ascribe M in such a way that the self-ascription of M is also an expression of the person's endorsement of M. To express your endorsement of M, as I said, requires the person to take M to be merited. But that one takes M to be merited does not entail that M will change when one changes one's assessment about M's merits. Take the example of a person with an irrational fear of spiders. There might be an occasion where he might come across a spider that he believes is a black widow. Believing it to be a black widow, he will take his fear for this spider to be merited; black widows are very dangerous and warrant fear. Because he takes his fear to

be merited, he has endorsed self-knowledge of such fear. But this does not mean that the fear is actually effectively judgment-sensitive. One might imagine that the person's partner, an arachnologist, as the ironies of life would have it, points out to him that the spider is not a black widow. But because his fear of spiders is an irrational phobia, recognizing that it is not a black widow will not alter his intense fear. The person can now judge that his fear is not merited, but he will be afraid nevertheless. His fear, over which he had endorsed self-knowledge, was not effectively judgment-sensitive.

The example shows that it is possible to have endorsed self-knowledge of recalcitrant mental states, i.e. of mental states that are not effectively judgment-sensitive. This is so because, as the example illustrates, it might just so happen that the mental state turns out to align, here and now, with the person's recognition of merits of the the mental warrant's.

A mental state that is effectively judgment-sensitive does not merely align with the current judgment of the person but would align with any potential future judgment that this person would make about the merits of its warrants. A state known with endorsed self-knowledge might just happen to be fixed and recalcitrant but accidentally aligned, in this particular case, with the person's judgment about the mental state's merits.

But even though endorsed self-knowledge of M does not necessarily entail that M is effectively judgment-sensitive, it is still the case that there is an internal connection between this form of self-knowledge and this type of mental state. If things are going well, endorsed self-knowledge will be of judgment-sensitive mental states. In general the person's recognition of M's merits will make it the case that the person holds M. Holding M depends on holding M to be merited. And when this happens, when holding the mental state depends on the recognition of its warrants, it will be effectively judgment-sensitive.

When a person has endorsed self-knowledge there is conceptual space to raise questions about M's merits, questions that are meant to transform the mental state if the person does not think that it is merited. That the conceptual space to raise this questions opens up does

not entail that these questions will necessarily emerge nor that these mental states will reflect the responses that one will provide to these questions. That this conceptual space opens up, however, does imply that these questions *can* emerge and that this mental state *should* be transformed in accordance with the person's response to these questions. Moreover, and as I have suggested earlier (and will further support below), if the mental state is not transformed in this way, it is because there has been some impediment or interference that did not allow the mental state to respond as it should to the person's assessment of the merits of holding it. When things are working well, when one has endorsed-self-knowledge of a mental state, such mental state is effectively judgment-sensitive.

Allow me now to discuss the relationship between mental states that are effectively judgment-sensitive and those that are known with endorsed self-knowledge. It might be natural to think that if a mental state is effectively judgment-sensitive, then this alone will put the person in a position to know it with endorsed self-knowledge. After all, for a mental state to be effectively judgment-sensitive *just is* for that mental state to respond to the person's judgments about the merits of the mental state's warrants. And it is natural to think that if a mental state responds to the person's judgments, then the person will know these state of mind with endorsed self-knowledge. But this argument, to work, requires that one assumes that the judgment that one is speaking about is a conscious judgment. And this is a premise that readers are unlikely to grant readily, particularly given that there are many judgment-sensitive mental states that are unconscious, but which become modified by different kinds of input from the world, most of it happening unconsciously.

To argue that an effective judgment-sensitive has to be known with endorsed self-knowledge one needs to argue, as I did in chapter 2, that properly responding to a judgment that is meant to be rational requires both that the judgment and the corresponding underlying mental states are conscious. And this suggests that for a mental state to be properly called "effectively judgment-sensitive" requires that the judgment and the mental state is conscious.

Being conscious of a mental state that is effectively judgment-sensitive, however, requires more than a mere capacity to formulate it in words, it requires the capacity to know it with endorsed self-knowledge.

Establishing an internal relationship between effectively judgment-sensitive mental states and mental states that are known with endorsed self-knowledge will be instrumental to show, below, that having these kinds of mental states is required to attain mental health.

Moran's articulates the notion of transparent self-knowledge in the service of explaining what epistemologists call "self-knowledge," i.e. the supposed privacy, intimacy and authority that we have over some of our mental states. My project is not attempting to do this and is meant to be neutral with respect to the theories of self-knowledge offered by epistemologists. I share with Moran the conviction that endorsed self-knowledge provides one with a particular way to know some of our mental states. My interest, however, is not to argue, as he does, that the alleged intimate and authoritative way in which we know our some of our mental states depends on this particular way to come to know them. My interests is in understanding why possessing the ability to relate with my self-knowledge in this way is ethically significant.

V.1.5 An Illustrative Example

I'd like to conclude this section by transcribing an example that Moran provides in his book and which will help to consolidate our understanding of transparency. It is an example to which Moran returns a few times in his book, which Lear uses to delineate his criticism, and to which I will refer back a few times.

> The person who feels anger at the dead parent for having abandoned her (...) may only know of this attitude through the eliciting and interpreting of evidence of various kinds. She might become thoroughly convinced, both from the constructions of the analyst, as well as from her own appreciation of the

evidence, that this attitude must indeed be attributed to her. And yet, at the same time, when she reflects on the world directed question itself, whether she has indeed been betrayed by this person, she may find that the answer is no or can't be settled one way or the other. So, transparency fails because she cannot learn of this attitude of hers by reflection on the object of that attitude. She can only learn of it in a fully theoretical manner, taking an empirical stance toward herself as a particular psychological subject.[11]

V.2 The Therapeutic Role of Judgment Responsiveness

V.2.1 Judgment-Responsiveness as a Therapeutic Aim

Moran emphasizes that the normal function of the first-person present tense of 'believe' is to declare one's view of how things are out there. This, he thinks, follows from the fact that to believe some proposition *just is* to believe that it is true.[12] Built into the idea of a belief is the fact that it purports to track the truth. To be a believer is to be someone whose understanding of the facts of the matter settle what she believes.[13] And this entails that to be a believer is to be bound by what he calls "the transparency condition," i.e. bound to be prepared to respond to questions about one's beliefs about X by considerations about the facts with which X is concerned, not about one's psychology.[14] This is a feature that applies not just to belief but to any judgment-sensitive mental state. After all, a judgment-sensitive mental state is precisely a state that could be formed and should be transformed by making

11. Richard Moran 2001, p. 85.

12. Richard Moran 2001, p. 74.

13. Richard Moran 2001, p. 76.

14. For versions of this type of argument in Moran see: Richard Moran 2001, pp. 68, 84.

judgments about the objects with which it is concerned. A judgment-sensitive emotion should be responsive to the person's judgments and deliberations about the fittingness of such an emotion. "Coming to believe that some fear of mine is unfounded will normally change my emotional state, replacing fear with something else, perhaps relief."[15]

The pressure to make our judgment-sensitive mental states *effectively* judgement-sensitive is a normative pressure that comes, not merely from the nature of such states, but also from our recognition of ourselves as holders of them. And because we recognize ourselves as creatures who hold them, the nature of these states, the fact that they belong to the genus "judgment-sensitive," places a normative constraint on us.[16] These states are ones that ought to be effectively judgement-sensitive. And because we recognize this fact, there is a normative pressure for us to have effectively judgement-sensitive mental state.[17]

Of course, human beings hold judgment-sensitive mental states that are *not* effectively judgement-sensitive. However, in so far as we are creatures who purport to hold judgment-

15. Richard Moran 2001, 54. See also, pp 58-59.

16. Cf. McGeer and Pettit 2002.

17. Justin D'arms and Daniel Jacobson argue that to say an emotion is appropriate, in the sense of being fitting (i.e. effectively judgment-sensitive), is different than to say it is morally appropriate: "there is a crucial distinction between the question of whether some emotion is the right way to feel, and whether that feeling gets it right" (D'Arms and Jacobson 2000b, p. 66). Putting D'arms and Jacobson's thesis together with what I have just argued, entails that being mentally healthy, in the sense just highlighted, can come apart from being moral.
The idea that mental health requires one to hold immoral mental states would be disturbing for most virtue ethicists. And although properly investigating how these virtue ethicist could respond to D'arms and Jacobson's challenge is beyond the limits of the chapter, I'd like to give a hint as to this challenge could be addressed. Even if one is willing to grant to D'arms and Jacobson that there are cases where it is inappropriate to hold an emotion that is, nevertheless, fitting (I will discuss some of these cases later in the chapter), if one is moved by the argument I just put forth and one wants to hold that morality and mental health do not conflict, then one is bound to believe that morality cannot, in general, demand us to hold mental states that are not fitting.
A virtue ethicist who granted this would be committing herself to a particular constraint on the kinds of mental states that morality can require us to hold. A supposedly moral demand would be shown not to be truly moral if it requires us to hold mental states that are not fitting, that do not respond to what the facts of the matter warrant. This constraint could require a revision the alleged morality or immorality of holding certain mental states (for instance, it would show that it is not immoral to feel certain types of envy, when these forms of envy are fitting). The revisionary extent of these changes, however, would be far less reaching than D'arms and Jacobson's paper suggests, something that would be shown by a careful and more textured analysis of the examples that they discuss and which are meant to illustrate that there are fitting but allegedly immoral emotions.

sensitive mental states, this is not how it *should* be. And because we have the capacity to recognize this, there is an internal pressure for us to seek that our judgement-sensitive mental states are *effectively* judgement-sensitive. Moran expresses this nicely: "'Taking my beliefs to be true' is not an empirical assumption of the sort that I might make with respect to another person. Rather it is a categorical idea that whatever is believed is believed as true."[18]

V.2.2 Judgment-Responsiveness Is Not the Only Therapeutic Aim

I am proposing here that having states of mind that are effectively judgement-sensitive is a central criterion to judge the mental health of a person. But this does not meant that it is the only criterion. The notion of mental health is complex and multifaceted. To be mentally healthy is not merely to be a rational believer. Mental health is inextricably linked to what is often called psychological health, which is determined by the person's capacity to live a good and flourishing life.And because a flourishing life is more than merely a rational life, there will be cases where a person's flourishing could be enhanced by relinquishing the aim that all of one's mental states are effectively judgement-sensitive.

It is not difficult to come up with examples where the person's ability to live well might require him to hold mental states that do not pass rational scrutiny, in particular beliefs which do not track the truth or emotions that are not warranted given the circumstances. Think, for instance, of a father for whom the death of his only daughter, a young woman killed in a car accident, has the potential to utterly paralyze him (and not merely during a reasonably mourning period, but for his whole life). It was a true accident; no one is to blame. Her friend was driving safely when a hinge unexpectedly broke, sending the car through a cliff. This person is such that if he does not find anyone to blame for her death, he will blame himself his anger will be self-directed and he will be stuck in a depression from

18. Richard Moran 2001, p. 77.

which he will not be able to emerge. If recognizing that this was a mere accident would be so devastating for him in the long run, it seems right to say that it is better for him *not* to track the truth and to find someone to blame. We might even say that it would be good for him if he was endowed with psychological mechanisms that protected him from confronting this truth, despite sufficient and convincing evidence being presented to the contrary. He might even be in a position to say: "although there is overwhelming evidence supporting that my daughter's death was not the fault of the driver, although it is unwarranted to resent and despise him, I nevertheless do." Some philosophers will argue that this is an irrational response. From the point of view of him as a holder of judgment-sensitive mental state this is entirely correct. But it is not an irrational response from the point of view of him as a human being attempting to live a flourishing life; insofar as this protects him from dramatically undermining his overall flourishing, anger, resentment and contempt for the driver are emotions that he *should* have, even though they undermine the normative constraints imposed on him as someone who holds them. (This possibility brings out clearly that being a flourishing human being is not merely being a proper holder of mental states. It also brings out that the question "what mental state *should* I hold?" brings with it more than one possible normative dimension.[19]

I just conceded that the normativity at stake in the question "how should I live?" outstrips the normativity that judgment-sensitive mental state require. When a person asks herself "what *should* I believe?," and this question is meant to provide her with part of the answer to "how should I live?," she is not merely asking about the normative constraints of judgment-sensitive mental states; she is rather asking about the way in which holding this or that mental state furthers or impedes her flourishing. I also conceded that there are cases where considerations about the expediency of a belief or emotion can override considerations about its rationality. But it is important to recognize that these cases, even if they might

19. This is a point that D'arms and Jacobson bring up nicely (D'Arms and Jacobson 2000b).

promote a person's health, are nevertheless deviations from the way in which the person should ideally live her life.

In the case of the father in our example, it seems clear that even though being self-deceived about the circumstances around his daughter's death is best for him, there is something wrong if he needs to be self-deceived about this to live a flourishing life. My natural inclination is to say that this father has a disturbed relationship with his daughter or the world, a pathological relationship that makes him incapable of facing the truth that someone can die by accident, without anyone's fault. A fully virtuous person is someone who can face and overcome the tragedies of life with courage, someone who would find the accident that killed his daughter's death painful and devastating, but who could nonetheless face it, eventually overcome it, and be able to live a flourishing life despite it.

V.2.3 Depressive Realism and Mental Health

There are other examples where failing to track the truth might also be conducive to a person's flourishing. It has been said that we are more creative in some of our inquiries if we evade acknowledging some of our limitations or the difficulties involved; fully acknowledging them might discourage us from undertaking the inquiries or sustaining them with vigor.[20]

Some psychologists have actually defended the thesis, grounded on the results from experiments on depressive realism, that biased illusions are a necessary part of mental health.[21] If the results and interpretations of the experiments in depressive realism are correct, positive biases and self-deception would be pervasive in what we consider to be healthy individuals. This, however, is insufficient to prove that these kinds of evasions are inescapable or that they display health at its best. As I will now argue, the proper and fully flourishing life

20. Many of the remarks of this section are indebted to Martin 1986, p. 126

21. Depressive realism is the thesis according to which depressive individuals have a more realistic assessment of reality than non-depressed individuals. For a review of the literature on depressive realism see: Dobson and Franche 1989. For a meta-analysis on the studies on depressive realism see: Moore and Fresco 2012.

is one where the person is able to pursue her projects with vigor *while* acknowledging her limitations. And this is so because, as I will contend, being an excellent holder of mental states is part of what is at stake in being an excellent human being.

Allow me to focus first on the case of beliefs. The question "how should one live?" is a question that only a self-conscious believer can pose and answer. Accepting this question as meaningful, accepting that reflecting on it has repercussions on how one actually should live, entails accepting, amongst other things, that one is a believer. To be bound by this question is, therefore, to be bound by the normative constraints of belief. I've said earlier that it is internal to holding a belief that it is meant to track the truth. This internal connection between belief and truth imposes a normative pressure for self-conscious believers; self-conscious believers are bound to believe what is true not what they wish. The argument, of course, generalizes to all judgment-sensitive mental states. Being bound by the question "how should one live?" means being bound by the normative constraints of judgment-sensitive mental states: these states should be rational and, in so far as our judgment capacity works well, effectively judgment-sensitive. Because we recognize ourselves as holding these states, and because we recognize their normative constraints, we are bound by them. Being bound by the question "how should one life" is, therefore, being bound by the normative constraints of judgment-sensitive mental states, constraints that we recognize and to which we attempt to conform. This is what justifies that living an excellent life involves having a mind whose judgment-sensitive mental states are rational and transparent.

It is relevant to underline the fact that although what I have been saying is compatible with the results that support depressive realism, it denies the phenomena. I am willing to concede that there are areas where depressed individuals can be more realistic than non-depressed individuals; the research provides some evidence for this. But this does not undermine the thesis that a healthy individual is someone capable of perceiving the world without distortion. It might be the case that most individuals usually considered healthy

have, in general, more positive biases or distortions than depressed individuals. But this does not prove that "biased illusions represent a necessary part of mental health." Health is a normative concept and there might well be aspects of what this norm involves that are not represented in the population in a statistically significant way. I argued that mental health involves holding rational and transparent judgment-sensitive mental states. The results that give support to depressive realism bring out that one dimension of health along which most people's health falls short has to do with the accuracy of their perceptions and judgments (most people, otherwise healthy, have positive biases and illusions). But this does not imply that to be a flourishing believer we need to become depressed. Being a moderately healthy individual who is self-deceived and biased or an accurate but depressive person are not the only two available alternatives on the table. There is also, of course, the possibility of being a non-depressed individual who is not self-deceived. And it is this individual who sets up the standards of health. Most people considered healthy might not be realistic and have positive bias; most people who are mildly depressed are more accurate in their judgments and assessments. But what this would prove is not that being healthy requires self-deception but that being a realistic non-depressed individual is harder than one would have initially supposed; it suggests that we are in need of cultivating hope to a much greater degree so that we are better able to deal with the pain, anxiety and potential depression that might be involved in having an accurate worldview.[22]

V.2.4 Judgment-Responsiveness Is a *Central* Therapeutic aim

So far I have argued that a person's rationality and her capacity to have effectively judgment-sensitive mental states can (and should) be relinquished when this is needed to further other life goals that impede this person's flourishing. I qualified this by contending that, when the person relinquishes her judgment-responsiveness to some states of mind in order to promote

22. This idea that hope is necessary to live a healthy flourishing life is further explored in: Martin 1986, pp. 126–131, Sherman 2014a, and Jolley 2014.

her flourishing, her life cannot be a proper exemplar of a flourishing life. Up to now I have only substantiated that a person's life should be bound by rationality and judgment-responsiveness. This only proves that rationality and judgment-responsiveness are part of mental health. My thesis, however, is stronger. What I want to argue is not merely that judgment-responsiveness and rationality are *a* part of mental health but that they are a *central* part of it. I will conclude this section by arguing for this stronger conclusion.

Often times self-deception is said to be problematic because it impugns certain important virtues such as honesty and sincerity. And one might think that what I have been saying suggests that sometimes these virtues should be sacrificed to develop other important values, like a subjective feeling of happiness. But although it is perfectly appropriate to characterize honesty and sincerity as values that are on a par with other values and that can compete with them,[23] it would be mistaken to see judgment-responsiveness or rationality as values that are on a par with these kinds of values. The aim of achieving mental states that are effectively judgment-sensitive is constitutive, not merely of having a mentally healthy life, but of having a mental life at all. Moran reminds us that "as an empirical matter, the fact of anyone's believing P leaves open the question of the truth of P itself. (...) But for the person herself, if her belief that it is raining does not constitute the question's being settled for her, then nothing else does. To have beliefs at all is for various questions to be settled in this way. (...) [T]o be a believer at all is to be committed to the truth of various propositions."[24]

Although we can make sense of the idea that prudential considerations might lead a person not to believe something that she finds unwarranted, this is a thought that we can hold only with difficulty. There are deep conceptual tensions in the thought that a person can deceive herself or that her will plays a role in determining her beliefs, tensions which

23. See, for instance, Williams 2002

24. Richard Moran 2001, p. 77.

have led many philosophers to reject altogether that there can be such a phenomenon as self-deception.

The life of a human being is an answer to the question "How should I live?"[25] Being able to answer this question (in fact any question at all) presupposes that the person's mental capacities are able to address it. One minimal and ineliminable aspect of this is the person's capacity to settle her beliefs by reflecting on the truth of the matter (and not on, say, prudential considerations). Rationality and judgment-responsiveness are at the heart of the capacity to respond to a question. They are, in fact, presupposed by the capacity to talk. And this means that rationality and judgment-responsiveness are not values, like honesty or generosity, from which a believer could walk away, they have logical precedence over such values; the possibility of such values conflicting depends on them.[26]

Even if we can make some sense of the idea of a believer endorsing a belief because of considerations other than the belief's warrants, and even if we can make some sense of self-deception, our ability to make sense of the phenomenon is limited. The occasions where the person holds judgment-sensitive mental states that are not effectively judgment-sensitive have to be isolated. If most of a person's judgment-sensitive states lack judgment-responsiveness, if most of them cannot be settled by the person observing the facts, it starts to become difficult to think of this person as someone who can hold judgment-sensitive mental states, to think of such states as being "judgment-sensitive" *at all*, and, as a consequence, it is difficult even to make sense of the idea that the person has a mind.[27]

25. With this I don't mean to suggest that every human being has explicitly reflected on this question or that her actions are deliberatively and self-consciously aimed to respond it. My claim is rather that the person's actions, emotions, etc. are always a response to it. Even if the question might not have been self-consciously addressed, the person's actions, the values they manifest, and the way in which these manifested values hang together, are all things that entitle one to ask and press her about what they reveal about her life and how she thinks she ought to live it.

26. Incidentally, I think that this is one of the considerations that has led philosophers to put rationality *at the heart* of what it is to be a human being, a view to which I could have appealed to support my thesis that judgment-responsiveness is *central* to human flourishing.

27. This is something that almost all the philosophers who advocate for seeing self-deception in a positive light concede. They all grant that self-deception can be "user friendly" only if it occurs locally, only if it is

Allow me to recapitulate what I have said in this section. I have argued that rationality and judgment-responsiveness are constitutive of our capacities to hold judgment-sensitive mental states and to talk and respond to questions. They are, therefore, constitutive of our ability to live a life which can be seen as a response to the question "how should I live?" As such, they have a logical priority over goals that are usually characterized as virtues. Although I conceded that there are occasions where a person might relinquish rationality and judgment-responsiveness to further other human goals that are important to her flourishing, she can only do this in a limited way if we are to make sense of the idea that she has a mind at all. Finally I argued that, even if this is something limited and local, relinquishing rationality and judgment-responsiveness should always be seen as a form of self-undermining. And this suggests that, although being a flourishing human being is not merely being a flourishing believer, the former is central to the latter.

V.3 Moran's and Lear's Therapeutic Ideals

V.3.1 Lear's Account of First-Person Authority

In his latest book, *A Case for Irony*, Lear proposes an alternative account of first-person authority to Moran's which is also meant to be faithful to Freud's practice. Lear criticizes Moran both for his account of self-knowledge and for his characterization of psychotherapy.

Lear's characterization of first-person authority is inspired by the expressivist account of David Finkelstein. According to this account, a person speaks about a state of mind with first-person authority when she can express this state of mind by self-ascribing it. Lear attempts to make this vivid by contrasting two cases. In the first case, I assert my anger based on the behavioral evidence. In stating that I am angry I am offering a true report based on the available evidence. But when I do this "I am not thereby *expressing* my anger.

not pervasive in the person's life.

For my anger is not *there*, present in the utterance."[28] This is a case where, according to Lear, I don't have first-person authority. The case where I *do* have first-person authority corresponds to a situation where I say "I'm furious with you!" in the midst of a boiling rage.

> In this case, the anger itself is present in the verbal self-ascription of anger that is directed at you. In this case, the self-ascription of anger is itself an *angry* expression. Unlike the former case, I do not have to observe myself to know that I am angry. I just *am angry*; and the form my anger at you takes on this occasion is the angry verbal self-ascription of anger directed at you. In this case, the utterance "I am *furious* with you!" may replace other forms of angry expression such as shaking with rage or impulsively striking out; or getting depressed or feeling guilty.[29]

This cursory look at Lear's characterization of first-person authority is sufficient to reveal that Lear's definition of first-person authority corresponds to my definition of first-personal self-knowledge. After all, a person has first-personal self-knowledge, precisely, when she can express a state of mind by self-ascribing it. It is worth pointing out that when things are going well, first-personal self-knowledge of a mental state involves both the capacity to express the mental state *and also* the capacity to express one's endorsement of it in a self-ascription. Lear, for reasons that will become clear soon, is particularly interested in cases where the second part of the conjunction does not hold, i.e. in cases where one has what I have called "merely-expressive self-knowledge," i.e. where the person can express her mental state in a self-ascription even if she *cannot* express her endorsement of such mental state.

The self-knowledge which Moran wants to vindicate, as I already mentioned, corresponds to what I have been calling endorsed self-knowledge. Its first-personal nature, however, goes beyond the mere capacity to express the mental state in a self-ascription. If a person has

28. Lear et al. 2011, p. 52.

29. Lear et al. 2011, p. 52.

endorsed self-knowledge, she is usually in a position to make it the case that she holds a mental state M by judging that it is merited to hold such mental state. Only the subject of a mental state can have this kind of authorial relationship with her mental states.

As a reminder of how these different concepts hang together I will replicate the tree that represents the relationship between these different types of self-knowledge on which I am focusing.

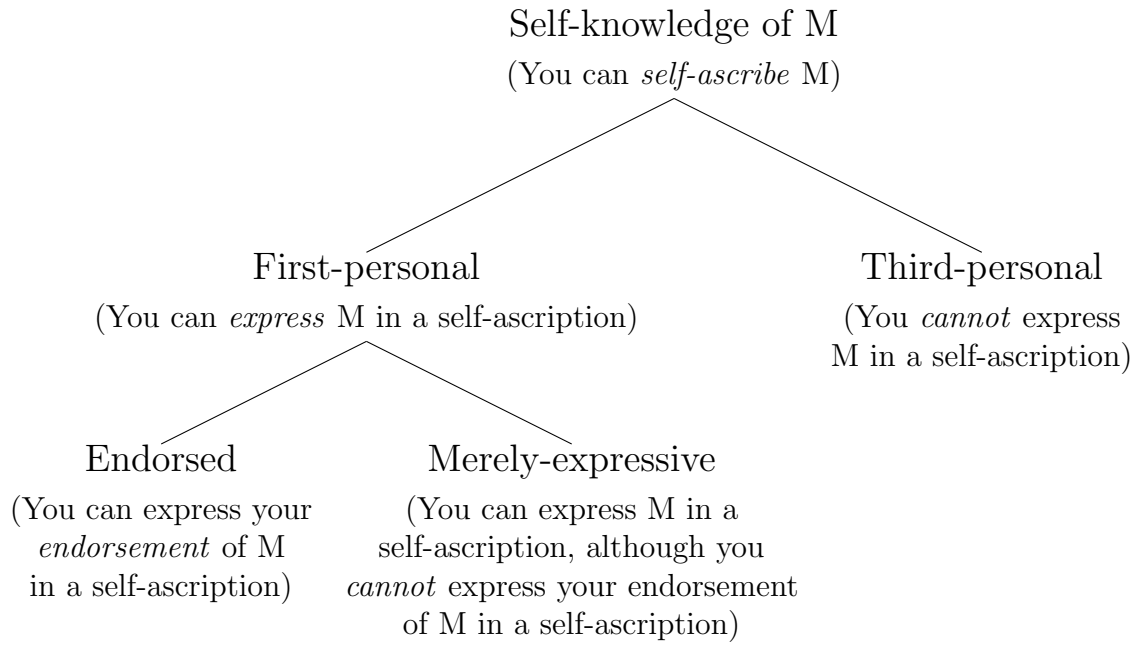

Figure V.2: A Taxonomy of Self-Knowledge

V.3.2 Moran's Either/Or

In his latest book, *A Case for Irony*, Lear criticizes Moran's account because, he claims, it offers an either/or: either self-knowledge is acquired through the eliciting and interpreting of behavioral evidence about the person or self-knowledge is transparent (*à la* Moran).[30] Throughout his book Moran usually presents only these two alternatives to characterize the way in which a person can know her mental states. Lear reminds us, however, that these

30. Lear et al. 2011, p. 121.

are not the only two possible alternatives in which a person can come to know herself. In elaborating this criticism of Moran, Lear states that a central part of the psychoanalytic practice consists in getting the patient to know her mental state with merely-expressive self-knowledge, getting her to express consciously her emotions, even if these are irrational and recalcitrant to her judgment about the merits of holding them, that is to say, even if they are *not* effectively judgment-sensitive. This form of acknowledgement is not captured by the two options that Moran usually presents to the reader; it is neither an avowal that involves that the mental state is effectively judgment-sensitive nor an acknowledgment acquired through the eliciting and interpreting of behavioral evidence.

To put it in the terminology I have been using, Lear's complaint is that Moran conceptualizes self-knowledge *either* as endorsed self-knowledge *or* as third-personal self-knowledge. In doing so, he does not leave space for merely-expressive self-knowledge.

Moran's somewhat idiosyncratic use of "avowal" contributes to the polarization of the conceptual space into the 'either/or' that Lear criticizes. Moran defines avowal as "a statement of one's belief which obeys the transparency condition."[31] This is not how the word is usually used. Within the literature on psychotherapy, in particular, "avowal" often refers to the emotionally charged acknowledgment of mental states that are, generally, not effectively judgment-sensitive. By restricting the meaning of "avowal" to convey merely *effectively judgment-sensitive* avowals, Moran rarefies the vocabulary, making it harder for a reader to find a way to talk about the form of self-knowledge that Lear discusses and to keep it in mind as a live alternative.

Lear's complaint about an 'either/or' in Moran's account is justified given the rhetoric of *Authority and Estrangement.* But it is important to say that there are a few of passages in the book where Moran explicitly concedes that a person can know her states of mind in

31. Richard Moran 2001, p. 101.

the way that Lear suggests.[32] And it is worth mentioning that there is an important reason why Moran does not discuss it as a self-standing third option: the form of self-knowledge about which Lear talks is not "first-personal" in Moran's sense. Moreover, from Moran's perspective, this form of self-knowledge is not structurally different from what he takes to be third personal self-knowledge. Moran lumps merely-expressive self-knowledge with third-personal self-knowledge. Neither of them obeys the transparency condition; neither of them is "first-personal" in his sense.

V.3.3 Moran Is Not Talking About Freud's First-Personal Self-Knowledge

When Lear argues that Moran's account presents an either/or, he adds that this either/or leaves no room for the psychoanalytic phenomena.[33] When Lear talks about "the psychoanalytic phenomena" he has in mind cases where the client is able to acknowledge her mental states with what one might naturally call first-person authority but which are not effectively judgment-sensitive. Take the case where the person knows that she is angry 'from the inside' despite the fact that she cannot avow her anger in Moran's sense: "She might feel her anger vividly, express it verbally, all the while knowing that—considered at the level of rational assessment—it is not warranted."[34] Lear argues that these types of states, which are not effectively judgment-sensitive, "lie at the heart of psychoanalytic work."[35]

If it is true that, despite his rhetoric, Moran is not committed to the either/or that Lear criticizes, then his account would leave space for such a phenomena. But leaving space for it is different from acknowledging its importance. I argued above that the aims that Moran identifies with psychoanalysis are at the core of the idea of mental health. Lear helps us see,

32. See, for instance, Richard Moran 2001, 32–3, 92, 101 and 131.

33. Lear et al. 2011, p. 121.

34. Lear et al. 2011, p. 70.

35. Lear et al. 2011, p. 56.

however, that Moran's account, which is supposed to illuminate the psychoanalytic practice, leaves out a really central part of what typically goes on in it.

I believe that there are strands in Freud's thinking that are in line with the therapeutic ideals of Moran. However it is important to keep in mind that whenever Freud distinguishes between two forms of self-knowledge that have a different therapeutic effect, he is usually referring to a form of self-knowledge similar to the one Lear has in mind, not to the one that Moran has. According to Freud, psychological impulses often become repressed because they clash with normative ideals that the person upholds. These repressed impulses crop up in disguised forms in the person's life and interfere with it. Freud argued that when these unconscious impulses are known by the person in the right way the symptoms vanish.[36] His work is full of examples of hysteric women whose symptoms are caused by their inability to accept and tolerate desires and emotions that were at odds with the conventional moral norms of their Victorian society, symptoms which vanished as soon as these patients came to recognize them. But Freud was clear that transforming these unconscious impulses into conscious ones did not make them effectively judgment-sensitive. In many cases, what the therapist does is get the patient to recognize that they have desires and emotions that are at odds with normative demands to which they are profoundly committed. Because these emotions and desires conflict with the patient's deeply held normative standards, they will not be endorsed by her, they will not, typically, think that they are warranted. And this means that the self-knowledge that the doctor is helping the patient to acquire is not effectively judgment-sensitive. The distinction between first and third-personal self-knowledge that Moran is drawing does not track the distinction that Freud was trying to establish.

36. Freud 1963, p. 281.

V.3.4 Moran's and Lear's Therapeutic Ideals

Moran's account of self-knowledge cannot illuminate a central dimension of the psychoanalytic practice. It also does not shed light on the distinction between the two forms of self-knowledge that Freud was putting forth and about which we are trying to get clear. It is important to recognize, however, that even though Moran's account does not provide the resources to characterize central areas of psychoanalysis, it does bring out a fundamental aim that has to be at the very core of any form of treatment that, like psychoanalysis, aims to promote mental health.

In his criticisms of Moran, Lear contrasts what he takes to be the therapeutic aims of psychoanalysis with the therapeutic aims that Moran puts forth. He writes:

> [T]he aim of psychoanalytic therapy is not to make one's fantasy [which includes the anger of the patient at her dead parent] *go away* even when it becomes conscious, which in effect is what is being called for when one states that the goal of analysis is that one's emotional life should obey conditions of transparency [*à la* Moran]. Rather, the goal is to find creative and life-fulfilling ways of living with fantasy [in the case of our example, life-fulfilling ways of living with her irrational but recalcitrant anger at her dead father]. (. . .) The success of therapy does not necessarily depend on the anger's *going away* in the light of rational assessment (transparency), but rather on whether one ceases to be *stuck* with the anger in rigid routines that one does not understand."[37]

Lear grants that there are occasions where the irrational emotions brought up in a psychoanalytic therapy abate and diminish over time.[38] But although he admits that this is valuable and therapeutic, he insists that this is not the aim of psychoanalytic therapy.[39]

37. Lear et al. 2011, p. 70.

38. Lear et al. 2011, p. 70.

39. Lear et al. 2011, p. 70.

Lear, then, acknowledges that there are two different strategies to treat judgment-sensitive mental states that are not effectively judgment-sensitive: transforming them so that they become effectively judgment-sensitive or learning to live with them despite the fact that they are not effectively judgment-sensitive.

These two outcomes are not on a par if we think that rationality is central to our nature as human beings and that the rational health of our minds is central to our mental health. A rational person is not just someone who is able to make judgments about the world; she is someone whose states of mind reflect these judgments. In so far as she recognizes herself as someone who is rational, she is bound to pursue states of mind that are effectively judgment-sensitive. And this means that it is a normative ideal that her mental states are aligned with her judgments. Of course, if an attitude is not responsive to rational assessment the person should find ways to live with it, as Lear suggests. But this is a coping mechanism that she should seek if she can't get these mental states to be responsive to rational assessment or if making them effectively judgment-sensitive is too painful, anxiety provoking or, speaking more generally, if it is less desirable given the consequences that judgment-responsiveness can bring to her overall flourishing. Earlier we granted that there can be occasions where self-deception is warranted within a person's life. My suggestion here is similar. Although judgment-sensitive mental states are meant to be effectively judgment-sensitive, there can be cases where it might be better for a person to cope with some mental states that are not effectively judgment-sensitive than to try to make them effectively judgment-sensitive. Attempting to make them effectively judgment-sensitive can take a huge toll in the overall well-being of the person and justifies a therapy that helps the person to cope with them instead of a therapy that aims to transform them.

In our earlier discussion I mentioned that holding on to illusions might be necessary for certain persons to flourish. I also said that persons who require these illusions are not being fully healthy nor fully virtuous; in so far as they are self-deceived their life cannot be said

to be fully flourishing. This is so because, as I argued, there is a normative pressure for us to get our judgment-sensitive mental states to be effectively judgment-sensitive.

Thus, the therapeutic aims laid out by Lear are a second best compared with those suggested by Moran. Therapists concerned with mental health should recognize that, at least in the ideal case, the aims laid out by Moran have a logical priority, as it where, to those laid out by Lear. It is only when seeking judgment-responsiveness fails, or when it is too costly to a person's flourishing, that a therapist should fall back and help the analysand cope and live with her non-judgment-responsive mental states.[40]

It might be worth to pause and bring out how the specific argument I am making here bears on wider issues within moral philosophy. The view I am defending is, at least in its general features, not too different from the family of views proposed by philosophers such as Christine Korsgaard, Angela Smith, Tim Scanlon, and, to some extent, Harry Frankfurt. According to the views put forth by these philosophers, we should aim to live a life that is unified by reason. The crucial words, here, are "reason" and "unification." In the next section I will argue that psychotherapies help us recognize certain nuances and complications with these views. In particular, that there are different ways in which learning to tolerate and accept our disunity is significant for mental health, ethical development, and, ultimately, our own unity. But I'd like to emphasize here that if the argument I just provided is correct, what psychotherapies have to teach us is not that the views of Korsgaard, Smith, and Scanlon are essentially flawed or mistaken, but rather that there are practical complications in implementing them. In saying this I am disagreeing with several authors inspired psychoanalysts such as Lear, Nancy Sherman, or Michael Lacewing who appeal to insights from psychoanalysis to criticize these views.[41] According to my view, what psychotherapies

40. As I have just said, a central premise supporting the conclusion that Moran's therapeutic ideals are superior to Lear's is that rationality is central to our nature and that our rational health is fundamental to our overall mental health. This is a position that is usually granted and one for which I will not argue here beyond what I already said in its defense earlier (V.2.4).

41. Lear et al. 2011; Sherman 2007; Lacewing 2008

can teach us about this family of views is that the ideal of unity that these philosophers promulgate cannot be pursued in any kind of way, not that such an ideal is misguided or mistaken.

V.4 Judgment-Responsiveness Can Get in the Way of Judgment-Responsiveness

I argued above that achieving mental states that are effectively judgment-sensitive is a central aim of therapeutic practice because the possession of effectively judgment-sensitive states is a sign of mental health. It is important to note, however, that seeking to have mental states that are effectively responsive to judgment should not be pursued at all costs. In this section I will actually argue, even if it might sound paradoxical, that there are different ways in which aiming at judgment-responsiveness can get in the way of achieving judgment-responsiveness. There are at least three ways in which this can be the case.

V.4.1 Accepting Emotions to Transform Emotions

Let me start by highlighting that when a person is too keen on making her mental states effectively judgment-sensitive she will have difficulty tolerating and accepting her conflicting mental states. When she rejects and disavows these mental states she might think that she has thereby eradicated them. Quite often, however, this does not address the conflict but rather hides it from view. As Sherman has said, discussing some of these issues, "[T]he conflict goes on, leaking through in neurosis."[42] Sherman mentions that "conflictual and concealed mental contents need a therapy of self-knowledge that does something other than continue to disavow them. They need to be heard from, in parliamentary fashion, and given

42. Sherman 1995, p. 232.

their own voice as a part of coming to be united with avowed and endorsed interests."[43] She adds that "giving them a voice, letting them speak through higher mental faculties, will help to resolve the conflict and disarm their subversive natures."[44] Sherman is describing here a phenomenon that is well accepted by psychotherapists, Lear included:[45] allowing a conflicting mental state to be expressed in words, allowing the patient to talk about it and explore it from the inside, often disarms it from its "subversive nature."[46] Transforming these types of states often requires that the person first accepts, experiences, and explores them. This is hindered if the person tries to confront them so as to examine the degree to which they are warranted. To fully explore and experience one such mental state the person needs, at least in the short term, to relax her aspirations of operating within the deliberative framework within which mental states are meant to be effectively judgment-sensitive.

This is the first case where aiming at judgment-responsiveness can get in the way of achieving judgment-responsiveness. Here, aiming at having mental state that are effectively judgment-sensitive needs to be held in suspension to allow the person to properly experience her conflicting mental states and to fully examine them from the inside. Experiencing and exploring these mental states is meant to render them more compliant to the person's conscious judgment, more compliant to her *qua* speaker. In so far as this happens, in so far as the mental states become effectively judgment-sensitive, the person can re-inhabit the deliberative standpoint. This is a case, then, where judgment-responsiveness is given up, albeit only temporarily, for the sake of judgment-responsiveness.

---

43. Sherman 1995, p. 230.

44. Sherman 1995, p. 233.

45. Lear et al. 2011, p. 70.

46. In the story that psychotherapists usually tell, which is also the story that Sherman is proposing, the changes in these mental changes are accompanied by changes in the judgments which conflict with them. As Sherman says: "[J]udgments of what is best may not remain the same once the disavowed is reclaimed and reinterpreted" (Sherman 1995, p. 233). I will return to this at the end of this section.

V.4.2 Giving up Judgment-Responsiveness Locally to Enhance Judgment-Responsiveness Globally

Of course, not all mental states are pliable in this way. Exploring and speaking about a conflicting mental state will not always transform it so that it conforms to the person's judgment. Lear reminds us that this is actually common. He appeals to his clinical experience to claim that there is a certain obduracy in the mental life of human beings. We have, he suggests, recalcitrant mental states which will, simply, not conform to the transparency condition, i.e. mental states that are not judgment-responsive. If this is right, then there is no point in trying to make these states of mind responsive to our judgment.[47] If there is a state of mind that is utterly recalcitrant, if we can be sure that it will be unaffected by deliberation or rational reflection, then what the person needs to do is, as Lear says, find life-fulfilling ways of living *with* it. This outcome, as I said earlier, is not as ideal as making such mental states effectively responsive to the person's deliberations and judgments. However, given the limitations imposed by its obdurate nature, it is the best outcome that she can hope for.

Aiming to make these obdurate mental states effectively responsive to judgment is, not merely useless, but often quite pernicious. People frequently seek therapy because they are unable to manage unwarranted emotions that they can't avoid having. These unwarranted emotions are usually accompanied by high levels of anxiety which are, in many cases, fueled precisely by the recognition that these emotions are unwarranted. That is to say, part of what makes clients so vulnerable to the anxiety that comes along with these emotions is, precisely, their lack of judgment-responsiveness. However, if these recalcitrant emotions are

47. There is quite a bit of evidence supporting that this is the case. Such evidence includes neurological facts about how old neural pathways are not eliminated but rather become circumvented, clinical observation from therapists dealing with this process in their consulting rooms, and empirical studies in psychology that have shown that many of the traits of the person are quite obdurate.

unalterable, if they will not become responsive to rational reflection, then fixating on trying to make them effectively responsive to judgment will not do any good. In fact, it will often make things worse. It contributes to making the person feel more inadequate about holding these emotions, and thus more anxious about her inability to make them respond to her self-conscious judgment.[48]

With these obdurate mental states, aiming at judgment-responsiveness often worsens the person's ability to live well with her mental state that are not judgment-responsive. In order to live well with mental states that are obdurate and recalcitrant to her self-conscious judgment the person needs to admit that she holds them, accept that they are at odds with her self-conscious judgments, stop worrying excessively about it, and cease to fight to change them. Being too fixated on achieving judgment-responsiveness will get in the way.

I began this section by talking about the case where the person gives up judgment-responsiveness to achieve judgment-responsiveness. In this second case we are supposing that the irrational mental state is obdurate and will not conform to the person's judgment. Thus, and unlike with the first case, the aim of making this mental state effectively judgment-sensitive has to be given up *permanently*. If we suppose, however, that when the person stops worrying about making these obdurate states of mind effectively judgment-sensitive her capacity to live with them improves, we come to see that even though the person is giving up judgment-responsiveness *permanently*, she is giving it up only locally. Supposing that the person's capacity to live with these obdurate mental states improves entails supposing that she is able to live a life that conforms better to her overall ideals about how she should live. And this suggests that, even though judgment-responsiveness is given up locally, it is given for the sake of enhancing it globally. How the person *actually* lives overall is now more responsive to her own considerations about how she *should* live. And thus, in these kind of cases, although she gives up her aim of making a local state of mind effectively

48. Acceptance Commitment Therapy is a recent but growingly popular therapeutic approach that makes quite a bit of this phenomenon (Hayes, Strosahl, and K. G. Wilson 2003).

judgment-sensitive, her life as a whole becomes more in line with the ideal that she has set up to achieve. She gives up judgment-responsiveness locally to regain it globally.[49]

V.4.3 Judgment-Responsiveness Can Entrench Our Defense Mechanisms

There is a further danger with being too intent on pursuing mental states that are effectively judgment-sensitive. As I will argue, one should aim to restore judgment-responsiveness *only* when the person's judging capacity functions well. When this is not the case, seeking mental states that are effectively judgment-sensitive do a disservice to the person.[50]

Freud states that by helping the patients know their unconscious impulses "the pathogenic conflict [was transformed] into a normal one for which it must be possible somehow to find a solution."[51] I suggested earlier that this pathogenic conflict was often caused by repressed

49. Moran uses transparency (i.e. what I have usually called judgment-responsiveness) to characterize a particular type of first-personal self-knowledge about individual mental states. I think that it is legitimate to apply this notion, not merely to individual mental states, but also to a life as a whole (in what I have been calling "global judgment-responsiveness"). It is certainly far-fetched to think that a person can have endorsed self-knowledge about her life; nobody comes to know how she lives merely by making judgments about how she should live. However there are aspects of a person's life which are effectively judgment-sensitive in this way, aspects of her life where she can say how she lives by stating how she should live. One of the central ideas at play in the notion of judgment-responsiveness and endorsed self-knowledge is that there is a interactive interaction between the person's judgments about what her mental states should be and her possession of those mental states. This dialectical structure, at the heart of the idea of judgment-responsiveness (and of endorsed self-knowledge), is also at work when we think about a person's life, and it is what justifies that we expand the notion of judgment-responsiveness in this way.
In this regards, it is worthwhile to note that a previous quotation, where Moran explains what is at stake in judgment-responsiveness, reads seamlessly if one replaces the references to mental states by references to the person's life:

> [the idea of judgment-responsiveness] lies in the requirement that I address myself to the question of what my life is in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the life of someone who happens to be me. This is not to say that one normally lives one's life through some explicit process of deliberation. Rather, what is essential is that there is logical room for such a question and that one's actual life is *answerable* to such considerations.

50. The danger I am highlighting has been discussed by other philosophers attracted by what psychoanalysis has to teach to philosophy. See, for instance, Lacewing 2014 or Sherman 1995.

51. Freud 1963, p. 435.

impulses that clashed with the person's normative ideals. Freud made it quite clear that there were a number of ways in which this conflict could be resolved but he suggested that the usual outcome was for patients to end up finding a middle path between giving in to these impulses and living according to the normative standards that conflicted with them.[52] This midway position entailed, for the analysand, realizing that there were important kernels of truth in the recalcitrant mental states that she had been characterizing as irrational. In tolerating and accepting these states the patient came to recognize that there where aspects of them which were justified. She came to recognize that her deliberations and judgments about her conflicting mental states were skewed, hindering her ability to recognize that there were aspects of them that were actually justified.

Take the case of the person who is angry at her deceased parent. We can imagine that her parent persistently neglected her. He was distant and seldom around when she needed him. When he died he sealed the fact that this would never change. And this was registered by the person in the form of anger at him for dying. We can tell a story about why her anger registered merely in this form, why the person did not register that, at bottom, her anger was caused by his persistent neglect, a story which might also account for the fact that her self-conscious judgment operated improperly, failing to register that being angry at his neglect was warranted. A therapist whose aim was merely to get the person to achieve self-knowledge that obeys the condition of transparency (i.e. endorsed self-knowledge) would help this person get rid of what she takes to be a conflicting and irrational emotion. This would be a very bad outcome; it would make her worse off. Before this pseudo-therapy she was at least registering the fact that her father was inadequate. After her anger has been excised, however, her defense mechanisms will have become entrenched and she will no longer be able to register a warranted anger toward her father.

Something important illustrated by this example is that although judgment-responsiveness

52. Freud 1963, p. 434.

is a necessary attribute of a healthy mind it is not a sufficient one. If the person's judging capacity is not working well, if, like in the example just discussed, her defense mechanisms interfere with her capacity to evaluate whether a mental state is warranted, achieving judgment-responsiveness will actually be a step backwards. In achieving judgment-responsiveness the person's conscious mind will become more unified: her mental state and her self-conscious judgment about it will be one. But this unity is a simulacrum of a rational mind. Prior to achieving judgment-responsiveness, she at least registered that her emotion was warranted. Even though now she appears more unified, this is at the cost of her failure to endorse something that she should endorse, to have an emotion that is warranted.

This phenomenon shows that the views about human identity held by philosophers like Frankfurt, Korsgaard, or Smith, need to be more nuanced. Characterizing the identity of a person merely by what she self-consciously endorses puts one in a situation where one is unable to distinguish cases where there is a mere simulacrum of unity from cases where the person's psyche is truly unified.

I am not suggesting that Moran's proposal is that one should seek judgment-responsiveness at all costs. I am quite sure that he would disapprove of the misguided therapeutic intervention just mentioned. He would forcefully deny that he is suggesting that the person's warranted emotional responses should be impaired to safeguard their judgment-responsiveness. He would likely respond to our examples by saying that, in the book, when he stated that the aim of therapy was to "restore to the person self-knowledge that obeys the condition of transparency" he was taking for granted that we were talking about a well-functional judging capacity whose judgments were rational.

But this assumption is very problematic. And this is so in two ways. First, it is an assumption that interferes with our ability to understand the way in which self-knowledge contributes to the self-transformation of human beings. Second, it leads us to conceive of our unconscious or recalcitrant emotions as irrational. Allow me to explain both, starting

with the first.

It is problematic to assume that one is talking about a well-functional judging capacity whose judgments were rational because this assumption interferes with our ability to understand the way in which self-knowledge contributes to the transformation of human beings.[53] If we are talking about actual human beings, we should not be too quick to assume that their judging capacities work well. Freud argued that human beings are plagued by defense mechanisms that interfere with their capacity to judge well. There has been a wealth of empirical material, provided by social psychologists and cognitive scientists, confirming what psychotherapists have long known: our minds are beset with mechanisms that interfere with our capacity to judge well.[54] Any account that seeks to explain how one is to restore mental health needs to keep this firmly in mind and cannot assume that our judging capacities work well. And this means that therapists should not be too quick to restore to the patient endorsed self-knowledge (i.e. what Moran calls "self-knowledge that obeys the condition of transparency"). This is a form of knowledge that should be restored only when the judgments involved are proper responses to the world.

The assumption that our judging capacities work well is also problematic in that it leads us to conceive of our unconscious or recalcitrant emotions as irrational. This perspective is pervasive in Moran's book. He seems to take for granted that when there is a conflict between our emotions and our self-conscious judgments, it is always the latter, but not the former, which are at fault. The example I just provided was meant to illustrate that this need not always be the case. There are cases where the unconscious emotions which clash with our self-conscious judgments *provide* the fitting response to the world. Arpaly's work,

53. It is not entirely clear to me whether Moran wants his account to shed light on how self-knowledge contributes to self-transformation and ethical development. There are moments, such as when he talks about psychotherapy and how it transforms people, where it seems that Moran is concerned with explaining how endorsed self-knowledge plays a role in the transformation of individual embodied human beings. For the most part, however, Moran seems to be interested in offering an account of what this process of transformation ought to achieve, an account of the aim of therapy and not of its process.

54. I discus some of this material in chapters II and III.

which I discussed in chapter II has additional examples illustrating case where a person has unconscious mental states that are more rational than the mental state to which she is consciously committed. In that chapter I argued, *pace* Moran, that there is a prima facie reason to think that unconscious mental states will tend to be less rational than conscious mental states, however I also conceded, *contra* Moran, that this need not always be the case.

These cases exemplify a third way in which judgment-responsiveness can get in the way of judgment-responsiveness. When Moran says that therapy should restore to the person her capacity to have effectively judgment-responsive mental states, he is not thinking of judgment-responsiveness as the mere alignment of the person's states of mind with her specific judgments about them. He is thinking of judgment-responsiveness as the alignment of the person's states of mind with what rationally warrants them. And this is something that can only happen if the conclusion of the person's judgments are actually warranted, that is to say, if her capacity to judge works well. When this is not the case one first needs to transform this judging capacity before one can aim to achieve judgment-responsiveness. This typically requires that the person recognizes that her deliberations have been skewed or misguided. And for this to happen she will often need to first learn to tolerate and accept mental states that she judges as irrational. She needs to be able to give a voice to these mental states so that this voice can engage the person's judging capacity and transform it. But because these mental states conflict with the person's self-conscious judgment, giving these conflicting states a voice will involve abandoning the deliberative standpoint temporarily. As the person's capacity to judge is transformed, as her judgments align with those emotions that are warranted, she can again re-inhabit the deliberative standpoint.[55]

This is a third case in which the right therapeutic intervention consists in putting on hold our immediate aspirations to attain judgment-responsiveness, and where this is done, ultimately, for the sake of attaining judgment-responsiveness.[56]

55. I will return to this point in the next chapter.

56. The example I used here is meant to bring out, in a way that was as clear as possible, that our uncon-

V.5 First-Personal Self-Knowledge and Their Role in Therapy

Lear believes that the mental states which are at the heart of psychoanalysis, and which the person is meant to come to know first personally, are obdurate and recalcitrant to her deliberation. These states of mind cannot be effectively judgment-sensitive and, thus, they cannot be known first-personally in the way that Moran suggests. As I suggested earlier, this means that the phenomenon which Lear could "first-personal self-knowledge" does not coincide with the phenomenon which Moran could call "first-personal self-knowledge."

Moran explicitly acknowledges that one can call these two forms of self-knowledge first-personal: "the first-person perspective, and the authority of the first-person, has two distinct aspects that are normally run together but can in principle come apart."[57] The first aspect is characterized by Moran as the epistemic immediate access which we have over the states of mind (with which Lear is concerned). The second aspect concerns Moran's transparent self-knowledge, the capacity for a person to know her states of mind by reflecting on whether she should hold them.

---

scious emotions might be right and our judgments about them wrong. It is worth mentioning, however, that the example is somewhat artificial. The story about a person's unconscious emotions and their corresponding judgments is usually more convoluted. Freud suggests that the usual outcome was for patients to end up in a midway position between giving in to their previously repressed impulses and living according to the normative standards that conflicted with them. I think that this is a typical outcome because there is often truth in *both* the unconscious emotion *and* the self-conscious judgment. Therapy entails a transformation in both. I presented my examples in a somewhat binary fashion to facilitate the exposition.
The story about how psychoanalysis transforms a person's mind is also more complicated than what my example might suggest. The point of analysis is not merely to transform one particular emotion and the person's judgments about it. It is meant to be much more holistic and encompassing, and it aims to transform recurrent patterns of thought, feelings, and actions. In the particular example I just discussed, the person undergoing a successful therapy will have, not merely developed a more healthy relationship with her anger towards her parent. She will also be able to recognize recurrent patterns in her interactions with figures of authority where she feels neglected, a feeling which will manifest in myriad ways that she will find hard to articulate. Psychoanalysis is meant to allow the patient to recognize how all these complex patterns play out, to transform them in so far as it possible, and to help her live with them in so far as she can't change them.

57. Richard Moran 2001, p. 91.

But although Moran concedes that there are these two aspects to first-personal self-knowledge, he also claims that the first is insufficient to characterize what we "ordinarily call first-personal self-knowledge."[58] As he writes: "From within a merely attributional awareness of herself, [a person who can express a mental state that she does not endorse by self-ascribing it] is no more in a position to *speak for* her feelings than she was before, for she admits no authority over them."[59] The authority about which Moran is speaking here is more robust that the epistemic authority involved in first-person authority. It is the authority that we have to make up our mind. When things are going well we are the authors of our mental states in a particular way: we can come to know that we hold a judgment-sensitive mental state by reflecting on its content and making up our mind about whether it is merited to hold such mental states. This capacity to make up our mind is distinctively first personal and is not available in the mental states which are at the heart of Lear's account.

One of Freud's central theses is that the person's cure requires that she comes to know the repressed impulses which are causing her symptoms with a form of self-knowledge that can be called first-personal. There are strands in Freud's thought according to which first-personal self-knowledge is the tool, the instrument, that brings about the cure. According to this line of thinking, self-knowledge is not the aim of treatment but merely a means to achieve the cure.

Moran's view contrasts sharply with Freud's. For Moran acquiring self-knowledge that obeys the condition of transparency *is* the aim of therapy not merely a means to achieve it. As I discussed earlier having endorsed self-knowledge is constitutive of mental health; a person's judgment-sensitive mental states are fully healthy only if she can know them with endorsed self-knowledge.

Lear's position is less easy to pin down. It is unclear to me whether Lear thinks that

58. Richard Moran 2001, p. 91.

59. Richard Moran 2001, p. 93.

merely-expressive first-personal self-knowledge is merely a means to "find creative and life-fulfilling ways of living with fantasy" or also a final end that should be attained for its own sake. Nevertheless, it seems quite clear that Lear thinks that merely-expressive self-knowledge is an ineliminable part of the *process* of successful analysis. For him it is not merely an aim that is pursued but also a means that helps to bring about the transformations that take place within the consulting room.

What I have just been saying should help one to see that the accounts of Lear and Moran should not be seen as competing proposals but rather as complementary. They each help us to understand different aspects of the self-transformation that psychoanalysis brings about, and the role that these two types of first-personal self-knowledge play in it. Moran's first-personal self-knowledge (i.e. endorsed self-knowledge) sets up the ideal that should be aimed at in therapy, while Lear's first-personal self-knowledge (i.e. merely-expressive self-knowledge) constitutes a mean that helps us attain this ideal.

An important question that this chapter opens up is why first-personal self-knowledge, of the type in which Lear is interested, is therapeutic. In particular how is it that giving voice to the mental states that conflict with our self-conscious judgments contributes to transform them into effectively judgment-sensitive mental states (perhaps through a joint transformation of the mental states themselves and of our judgments about them) or improves our capacity to live well with them, to be less stuck "in rigid routines that we do not understand." This will be the question that I will address in future work. The work we have done so far, however, should have helped us see that Moran's account of first-personal self-knowledge seems, on its own, ill-suited to address it. Acquiring endorsed self-knowledge is not so much a tool that paves the way for mental health as much as a necessary condition for it, a criteria which is required to establish that one has attained it.[60]

60. This chapter is indebted to the insightful comments of Stina Bäckström, Matthew Boyle, Noah Chaffets, Tupac Cruz, David Finkelstein, Mark Hopwood, Dhananjay Jagannathan, Nic Koziolek, Martha Nussbaum, Dasha Polzic, Daniel Rodriguez-Navas, Nancy Sherman, and Stephen Shortt.

V.5.1 Two Types of Subjects

The two types of self-knowledge that we have been discussing in this chapter are first-personal. For both it is true that they characterize a way of knowing a mental state M that is only available to the subject of M. I want to propose that one of the things that distinguishes these two forms of first-personal self-knowledge is a different conception of "subject." Getting clear on this will help us understand better a number of things. First, it will help to explain why we characterize our recalcitrant mental states as external; second, it will allow us to sharpen our understanding about the way in which these two ways of acquiring self-knowledge are first-personal; and third, it will show us that these different ways of acquiring self-knowledge are complementary proposals about different aspects of our being subjects of a mental life.

Within Lear's framework, a subject has first-personal self-knowledge when she can *express* a mental state by self-ascribing it. According to this conception, being a subject that can know her mental states first-personally just requires having an ability to express those mental states in a self-ascription. Thus, within this framework, "the subject" can know first-personally, not only judgment-responsive mental states but also mental state that are not judgment-responsive, in fact, even mental states that are not judgment-sensitive like pains or sensations.

Within Moran's account, a subject has first-personal self-knowledge about a mental state when she has the capacity to come to know it by judging whether it is merited to hold it. In this case, what "the subject" can know first-personally is much more circumscribed. Sensations, recalcitrant emotions, and, in general, mental states that are not effectively judgment-sensitive cannot be known by the subject when we understand "subject" in this more restricted way. Within this framework, "the subject" can only know from the inside those mental states which she can express in a self-ascription that is also an expression of her endorsement of the mental state.

Some readers will find it more natural to think about the nature of a subject and what is it for her to know her mind first-personally according to the model with which Lear is working. One might wonder whether there are any warrants to put forth a conception of the subject's capacities to know her mind that are as circumscribed as they are in Moran's framework. I think that there are. If we think that rationality is central to our being human it becomes plausible to think that these circumscribed capacities to know our minds are *particularly* distinctive of who we are. In so far as we think of ourselves as rational subjects, whose self-conscious rational judgments make a difference to what their mental states are, we are thinking of ourselves as subjects in this more restricted sense. It is also because we think of ourselves in this way, because we think that what we judge and deliberate self-consciously is at the center of who we are, that we are inclined to identify ourselves with those mental states that are the product of our rational activity, and to speak about our recalcitrant emotions as external or alien.

This inclination is what has lead philosophers like Harry Frankfurt, Angela Smith and Christine Korsgaard to go too far with this identification. These authors cash out our identity as persons in terms of what we endorse or decide. But, as we have discussed in other chapters of the dissertation, this conception is mistaken. Even though we often self-consciously describe those recalcitrant mental states that manifest in our behavior as external, taking too literally the idea that they are external makes us loose from sight that these recalcitrant states often play an important place in our identity, even if we might not endorse them. As I said earlier, although we might characterize these states of mind as external, they form part of an explanatory nexus which expresses and explains our behavior.[61]

Recognizing that there is something 'third-personal' about the immediate self-knowledge that we have of our recalcitrant states of mind is important to properly understand both types of first-personal self-knowledge. I mentioned previously in the chapter that, when

61. This, of course, is a merely promissory note about how an argument against this family of views would go.

Moran describes the epistemically immediate self-knowledge of the recalcitrant states of mind in which Lear is interested, he lumps them together with what I have been calling third-personal self-knowledge. Moran explicitly states that if a person lacks the capacity to endorse a mental state, if her judgments make no difference to what she beliefs, intends, or desires, her (perhaps epistemically immediate) knowledge about herself "may as well be about some other person, or about the voices in her head."[62] If a person can only relate to a judgment-sensitive mental state of hers in this way (regardless of how immediate it is) she will have "a kind of outsider's perspective on her attitude."[63] What Moran says here aligns with some of our common intuitions; we feel alienated from our recalcitrant states of mind; we disavow them because we consider them to be out of our control; in fact, these mental states might sometimes feel so external that we might say that it is as though they belonged to someone else.[64]

I suggested that reflecting on the type of subject that is at the heart of these two types of first-personal self-knowledge would help us to understand better that each of these forms of self-knowledge are not alternative proposals to characterize the same phenomena but rather complementary proposals about different aspects of our being "subjects of a mental life." Within Lear's conception of first-personal self-knowledge, to be a subject of a mental life is to be able to merely express it by self-ascribing it. Because the subject can express mental states that she deems irrational, this subject can come to know immediately states

62. Richard Moran 2001, p. 93.

63. Richard Moran 2001, p. 32.

64. If we are to be charitable to Moran's account, we need to be careful about how we understand the idea that these recalcitrant mental states are "external." Even though we might say that these mental states "may as well be about some other person," this is usually no more than a mere way of speaking. In so far as these are mental states that manifest themselves in our behavior and actions, they are very much our own, even if they are recalcitrant. And they are our own not merely because we, and only we, have over them what is usually called first-person authority (Finkelstein 2003, pp. 164-5). They are also our own because they are constitutive of our mental life. The recalcitrant beliefs, emotions and intentions that manifest in our behavior are not inert or innocuous. They express themselves in our actions and form part of the explanatory nexus which explains them. As such, they help to constitute who we are.

of mind that are not transparent in Moran's sense. This, of course, is not available to what constitutes the subject of a mental life within Moran's account. To be a subject of a mental life within Moran's account is to be someone who can come to know immediately her states of mind by deliberating and judging about the merits of holding them. This aspect of our subjecthood is at the center of our nature as rational beings. As such, it is warranted to think about our subjecthood in this restricted sense. Moreover, and as Mathew Boyle forcefully argues in "Two Kinds of Self-Knowledge," Moran's first-personal self-knowledge is in an important sense fundamental, because it is intimately connected with the very capacity for rational reflection and with our ability to use the first-person pronoun." A capacity that Boyle compellingly argues is presupposed by our capacity to express our recalcitrant states of mind by self-ascribing them.[65]

Proposing that Lear's and Moran's accounts of first-personal self-knowledge are complementary, and pointing the reader to Boyle's article, might suggest that I endorse Boyle's thesis that a satisfactory account of what epistemologists call self-knowledge cannot be uniform, that one cannot aim to provide a single account that explains all cases of "first-person authority." This is not the case; my proposal is neutral on this issue. What I have defended in this chapter sets up some restrictions on what a uniform account must look like, but it does not exclude that there can be such a unified account. If the story I am telling here is correct then it follows that a uniform account of first-person authority would need to Mark the fact that there is something distinctively first-personal about Moran's first-personal self-knowledge (i.e. endorsed self-knowledge) and something defective or alienated about Lear's (i.e. merely-expressive self-knowledge).[66]

---

65. Boyle 2009, p. 133

66. Allow me to sketch how this could work withing Finkelstein's account, an account which aims to be uniform. Finkelstein characterizes first-personal self-knowledge as the knowledge that you have over a mental state that you are able to *express* by self-ascribing it. His account would be compatible with what I am saying here if: 1) it cashed out the idea of "expression" in such a way that it would bring out that effectively judgment-sensitive mental states would characterize the proper functioning of this expressive capacity; 2) It accounted for the fact that when I come to know a recalcitrant mental state in which Lear is interested, I

am exercising this expressive capacity in a way that is somehow defective. Although it is beyond the scope of my project to try to offer such an account, one could appeal to the fact that rationality is central to who we are to secure the explanation of how the former and the latter are proper and improper exercises of our expressive capacities.

VI

Epilogue: Future Directions

In the previous chapter I argued that endorsed self-knowledge and merely-expressive self-knowledge have complementary roles in ethical development. Endorsed self-knowledge provides the unity of the person as a reasoner, speaker and doer. Endorsed self-knowledge is an aim that the ethical pilgrim aspires to achieve because endorsed self-knowledge is a necessary part of the person's mental health. I argued, however, that endorsed self-knowledge was not the kind of self-knowledge that is at play in psychoanalysis. Within therapy one deals mostly with mental states that are not formed and are not transformed by the person's own judgments about the merits of holding them. Therapeutic improvement depends on developing the capacity for merely-expressive self-knowledge. There are a number of questions that the previous chapter opened and which I intend to address in future work. The most important ones are: 1) Why is it important to come to know our recalcitrant irrational states of mind with merely-expressive self-knowledge? and 2) How does merely-expressive self-knowledge contribute to the development of endorsed self-knowledge (and thereby of mental health)?

We need to respond to these questions if we want to be in a position to properly understand the role of first-personal self-knowledge plays in ethical development. My dissertation focused mostly on explaining the importance of endorsed self-knowledge. I did so by ap-

pealing to the power of this form of knowledge to make us into more rational creatures. Understanding the role of merely-expressive self-knowledge in this process is important to qualify this account because it explains the importance of coming to know the parts of ourselves that are less rational in a first-personal way.

Theorists of psychotherapy do not agree on how therapy works. They have different accounts of how merely-expressive self-knowledge contributes to therapeutic improvement. Engaging with this literature has led me to the hypothesis that there is no single explanation that accounts for the role of merely-expressive self-knowledge in ethical development. Many of these explanations line up well with results in cognitive science and neuroscience.

It is the purpose of this epilogue to give a preliminary sketch of some of these explanations. Although my focus will be on merely-expressive self-knowledge, these explanations will bring out different ways in which all the varieties of self-knowledge that I have discussed in the dissertation interact and support one another. I want to argue that the ethical pilgrim's aspirations to know herself will lead her to acquire progressively higher forms of self-knowledge. If things are working well, third-personal self-knowledge will lead to merely-expressive self-knowledge and merely-expressive self-knowledge will, in turn, lead to endorsed self-knowledge.

VI.1 A divided person

In the introduction to the dissertation I mentioned that the dissertation's protagonist was an ethical pilgrim, a person who is not yet virtuous but who is trying to become virtuous. The ethical pilgrim aspires to think, feel and do certain things, but she often ends up being overcome by emotions, desires and beliefs that she deems unwarranted and which interfere with her aspirations. Take the person who, wanting to be patient with her kids, systematically loses her temper about minutiae, or a mirror image of Huck Finn, i.e. a person

who does not want to be racist but who is nevertheless racist.

The sketch that I will provide in this epilogue will focus on these kinds of cases, cases where the person holds mental states that she deems unwarranted and which interfere with what she takes to be her aspirations to be virtuous. I will refer to these kinds of mental states as "*conflicting*" mental states. To say that a mental state is "conflicting" is a shorthand for saying that it "conflicts with the person's judgments about the merits of holding such a mental state."

I said in the opening of the dissertation that the idea of an ethical aspirant, familiar to all of us, poses conceptual difficulties. An aspirant is positioned in two simultaneous but inconsistent perspectives that I called earlier the "ideal self" and "the actual self" (I.1.3). She is trying to live the life of the virtuous person, a life that requires having virtuous states of mind. But to determine what such virtuous mental states are this person will need to be able to embody the perspective of the virtuous person; it seems to require the person to already be virtuous. At the same time, because the ethical pilgrim is not yet virtuous, she embodies a position that deviates from virtue (a position which characterizes what I called her actual self), in the case of our parent a position of impatience, in the case of our organizer a position of envy.

In future work I intend to argue that part of the solution to this difficulty comes from recognizing that we are multifaceted. We have the ability to recognize that certain mental states interfere with our capacity to live virtuously. But this ability does not necessarily trickle down to fully inform these mental states. The fact that we can hold recalcitrant mental states reveals this kind of division. I believe that dual-process theories of cognition and the clinical experience of psychotherapies provide two complementary venues to investigate this problem and to articulate an account of it that is empirically informed. I also believe, and I will give some hints in this epilogue about why this might be the case, that the two forms of first-personal self-knowledge that I have discussed help us to see the tensions between the

ideal self and the actual self and the possibilities of resolution of it.

VI.2 Merely-Expressive Self-Knowledge and Our Sensibility

An influential series of experiments performed by Wilson and his colleagues consisted in asking participants to evaluate how their romantic relationships were going. Some participants had to perform these evaluations after discussing the reasons that justified their evaluations. Other participants were instructed to merely evaluate these things based on their gut feeling without providing any reasons. In these experiments the people who appealed to their gut feeling made better assessments: "the feelings people report after analyzing reasons are often incorrect, in the sense that they lead to decisions that people later regret, do not predict their later behavior very well, and correspond poorly with the opinion of experts."[1]

These results, I believe, have an important lesson to teach us about how to examine ourselves which I would like to articulate in future work. At bottom I intend to argue that there are many emotions that cannot be characterized as being entirely judgment-sensitive. In attempting to have endorse self-knowledge of this emotion one might end up, so to speak, overdoing it, attempting to provide justifications for things that should not be justified.

Love is a good example of this. Arguably, there are aspects of love that are not meant to be responsive to reasons. We do not fall in love with a person merely because he is a good catch. A person might have all the virtues that we expect to get from a partner and still fail to inspire us to love him. There is an important aspect of love that has to do with our subjectivity. When we embody a deliberative perspective and try to assess whether it is merited to hold this emotion by seeking reasons to justify it, we are doing violence to the emotion, imposing on it particular features that it does not have. There is

1. Timothy D. Wilson 2002, p. 170.

much in love that has to do with our sensitivity. And it is a danger to try to know it with endorsed knowledge because this form of self-knowledge tends to seek objective reasons and justifications to ground what one takes to be a merited mental state. Seeking to know these aspects of love with merely-expressive self-knowledge will keep the person from rationalizing what should not be rationalized. This is true about many of our most important decisions in life. Choosing a profession or a partner are all decisions that are not and should not simply be justified with the kinds of objective reasons that endorsed self-examination usually brings up.

Following Wilson, we can say that "[t]he story people construct on the basis of their reasons analysis (...) can misrepresent how they really feel."[2] And this is an unfortunate consequence that the search for endorsed self-knowledge might bring about. Developing merely-expressive self-knowledge is valuable to counteract these tendencies. Merely-expressive self-knowledge will bring us into contact with this subjective aspect of ourselves.

Research on dual-process theories and on the ways in which we have to access such process suggests that endorsed self-examination will tend to produce reports that reflect what Wilson calls "the conscious self" while merely-expressive self-examination will tend to produce reports that reflect what Wilson calls "the unconscious self."[3] Research surveyed in some of the previous chapters suggests that flourishing involves unifying these two selves, unifying your explicit and implicit attitudes. Merely-expressive self-knowledge will facilitate the process where these implicit attitudes come to the fore. Aspiring to have endorsed self-knowledge tends to disregard these implicit attitudes in light of the person's explicit attitudes. This, then is one of the ways in which expressive self-knowledge contributes to ethical development.

2. Timothy D. Wilson 2002, p. 170.

3. See, for instance, Gawronski, Hofmann, and Wilbur 2006; Gawronski and Bodenhausen 2012; Gawronski and Creighton 2013; Baumeister and Bargh 2014

VI.3 Controlling Mental States and the Varieties of Self-Knowledge

In what follows I want to provide a sketch about how I believe the different varieties of self-knowledge that I have discussed in the dissertation help us to control our conflicting mental states, a reflection that will help us understand better the role that merely-expressive self-knowledge plays in ethical development.

Let me preface what I will be saying by laying out some "brute data" about the role that self-knowledge (which involves putting mental states in words) plays in ethical development. A robust body of research in psychology has shown that "when people transform their feelings and thoughts about personally upsetting experiences into language, their physical and mental health often improve."[4] Pennebaker, a very influential scholar working in this topic, highlights that "[t]he mere [non-verbal] emotional expression of a trauma is not sufficient. Health gains appear to require translating experiences into language"[5] Experience from clinical practices as well as from what can be called 'practices on the self' have also led philosophers to argue that the unconscious nature of certain mental states explains a large part of its destructive power. As Martha Nussbaum claims: "Awareness is already progress towards cure."[6]

VI.3.1 Circumventing Conflicting Mental States

The sheer act of putting a mental state in language (be it first- or third-personally) allows the ethical pilgrim to examine it critically. She can articulate and investigate the way in which this mental state hangs together with her ethical values, to assess the conclusions that

4. Pennebaker and Chung 2014, p. 3.

5. Pennebaker and Chung 2014, p. 15.

6. Nussbaum 1994, p. 198. See also, Sherman 1995, 2007; Martin 2006; Lacewing 2008, 2014.

it entails and the premises that are meant to ground it. This allows the ethical pilgrim to properly evaluate whether she should hold it or not.

The ethical pilgrim's ability to put mental states into words also allows her to make plans to respond to it. In particular, she can come up with ways to prevent this mental state from interfering with her virtuous aspirations. Take a mirror image of Huck Finn. This person might come to know that she has racist attitudes. Given her non-racist aspirations, her knowledge of this fact might lead her to implement policies which can help her circumvent her racist beliefs or preclude these beliefs from being expressed. For instance, if she is responsible for hiring employees she might implement affirmative action policies or blind hiring practices. She might also fine-tune her workplace in such a way that she is faced with less situations or contexts which prime the activation of her racists beliefs.

These kind of responses to her mental states only require the ethical pilgrim to have third-personal self-knowledge. And it is certainly preferable for the ethical pilgrim to have this kind of self-knowledge than not to have it. But, as I have suggested in earlier chapters, this way to interact with her mental states is insufficient to secure proper ethical development. And this is so, among other things, because knowing a mental state in this way allows that the ethical pilgrim, at most, to circumvent it. And circumventing it is unlikely to be effective to change the mental state, something which, according to many accounts of virtue, the person aspiring to be virtuous would ultimately want. This form of self-knowledge allows the ethical pilgrim to react to her own mental state, but it does not allow her to transform it. She deals with this mental state as one deals, to follow one of Moran's analogies, with an obstacle in her path.

Despite its limitations, this is the conception of self-knowledge with which most scholars seem to work with. Social psychologists, in particular, are almost always guided exclusively by it. Their suggestions for ethical improvement involve inviting their readers, not so much to change themselves, as to take care that their environment where they live is such that it

is conducive to promoting the kind of behavior which they want to foster. This, however, is a very limited conception of the power of self-knowledge, one that arises from an utterly contemplative conception that thinks of self-knowledge as the mere acquisition of information about oneself.

VI.3.2 Clarifying Conflicting Mental States

There is a second way in which formulating mental states in language helps the ethical pilgrim control it. This form of control, however, requires first-personal self-knowledge. In her famous introduction to the work of Melanie Klein, Hanna Segal highlights the internal relationship between being able to name a mental state and the person's capacity to control it. While discussing the case study of Ann, a little girl, she writes: "[T]he real help I was able to give her was in naming the different feelings inside her helping her to know them, to differentiate them and, therefore, to feel more able to control them."[7] The control of which Segal is writing here is not the control that I discussed above. Being able to label a conflicting emotion is not merely valuable to help us implement policies to prevent its activation or manifestation. Labeling the emotion actually entails a transformation of the emotion itself, a transformation that is able to reduce, as Nussbaum suggests, the emotion's "destructive power."

This is so because conveying these mental states in language is also, in many cases, an act of informing and shaping these mental states. Pennebaker, who has strong research program investigating the effects of writing or speaking about trauma, has argued that "verbally labeling an emotion may itself influence the emotional experience."[8] He points to research by Norbert Schwarz which suggests that "defining and making attributions for internal feelings can affect the feelings themselves."[9]

7. Segal 1974.

8. Pennebaker and Chung 2014, p. 16.

9. Pennebaker and Chung 2014, p. 16.

Many of our mental states are not fully fledged; they are often inchoate and imprecise. The process of coming up with words to express them in language and to articulate their warrants is a process of informing and shaping the mental states themselves. By speaking and reflecting about an inchoate emotion, for instance, the pilgrim might be able to articulate and shape it. In this respect Nancy Sherman, a philosopher trained in psychoanalysis, has argued that "emotions can change and develop through continuing clarification of the stories that inform them. (. . .) [O]ur tendencies to feel anger or fear or shame can shift as we make more explicit to ourselves just what the beliefs are that rationalize those emotions."[10]

For this to happen, however, there has to be an tight connection between the mental state and its description, a connection that secures that the words will transform the mental state as it is being described. If the description is a mere referent to the mental state, like it is when one has third-personal self-knowledge, expressing it verbally is unlikely to make a difference in the mental states' nature. For the verbal description of a mental state to be effective informing it, the mental states needs to be "present in its verbal self-ascription."[11] It has to be, that is to say, known first-personally. Arguably, being able to have merely-expressive self-knowledge will be sufficient for this, because fleshing out the emotion merely requires that we are capable of putting the mental state in words, not that the mental state responds to our judgment about the merits of holding it.

VI.3.3 Transforming Mental States

When the mental state responds to our judgment about the merits of holding it we have a further way in which to control it. Although this is not a way that is available to transform the mental states in which this epilogue is meant to focus, i.e. conflicting mental states, it is worth highlighting it to show the way in which different forms of self-knowledge of our

10. Sherman 1995, p. 233.

11. Lear et al. 2011, p. 52.

mental states entail different kinds of control over them.

Above I spoke about how describing a mental state serves to determine it, to shape it from an inchoate feeling into a fully fledged mental state. In explaining this I also mentioned that developing our capacity for merely-expressive self-knowledge of a conflicting mental state will, typically, bring about a certain ability to control it, a capacity that does not merely amount to the devising of policies to circumvent its potential influence.

But the ethical pilgrim will typically want more than to merely control her conflicting mental states. She will likely want to be able to shape them in accordance with what she judges, i.e. to be able to make them not conflicting with her judgment. Having endorsed self-knowledge will, typically, secure this. As I explained in previous chapters, possessing endorsed self-knowledge of a mental state puts the person in a position where her evaluation of the merits of holding it will be typically reflected on her effectively holding such a mental state.

VI.4 Transforming Recalcitrant Mental States

Among the aims of this epilogue is to sketch some of the ways in which merely-expressive self-knowledge contributes to ethical development. The main idea that I want to put forth, and which I would like to develop further in future work, is that the three modes of engagement with our mental states sketched also explain the way in which the varieties of self-knowledge interact and contribute to mental health (and, thereby, to ethical improvement).

Let me start by mentioning how third-personal self-knowledge of our recalcitrant mental states allows for the development of merely-expressive self-knowledge. As I mentioned above (IV.8) there are cases where we are unable to know our conflicting mental states first-personally because there are emotional barriers that interfere with our capacity to know

them. The anxiety or psychic pain involved in knowing them might interfere with our capacity to know it first-personally. Coming to know these mental states third-personally is often a first step to tolerating the fact that we have them. As we come to have third-personal self-knowledge of these anxiety-producing mental states, our tolerance to acknowledge them often increases, and this allows us to gradually be able to come know these mental states with merely-expressive self-knowledge.

I hope to argue that acquiring merely-expressive self-knowledge of a mental state M is a step towards acquiring endorsed self-knowledge of M. To illuminate this, I plan to argue that mental states, particularly when they are connected with things that we value deeply, might appear recalcitrant in the short term but might not be recalcitrant when we take a longer term horizon.[12] To say that a mental state is recalcitrant is to say that, here and now, it is not effectively judgment-sensitive. But if the experience within the consulting room is to be a guide, it is safe to say that engaging our recalcitrant mental states with our self-conscious judgment will gradually lead us to have a more unified mental life where such mental states become amenable to respond to our judgments about the merits of holding them.

As far as I can see, the explanations for a mental state's recalcitrance can be reduced, mainly, to two. First, its recalcitrance might be explained by the fact that changing the mental state is emotionally difficult. Perhaps holding on to it has proven to be adaptive or the idea of letting it go provokes too much anxiety. Second, changing the mental state is cognitively expensive. It might require that a whole lot of other mental states are updated and transformed. At a neurological level, it might require rewiring firmly established neural paths that take time to rewire.

12. There are in fact neurological reasons that explain why it takes time to change our mental states: the neural circuitry which underlies such states needs to be rewired and this takes time.

VI.4.1 Dialoguing With Yourself

I want to propose (and I'd like to articulate this further) that we think of this slow process as a form of dialogue with oneself. I want to propose that the two parties that take part in this dialogue represent two conflicting perspectives that a single person takes on an issue: one party corresponds to the part of the person that is oriented towards and by the recalcitrant mental state M, the other part is oriented by a self-conscious judgment that finds that it is not merited to hold such mental state.

The first thing that I wold like to argue is that such dialogue requires that one has first-personal self-knowledge. In fact, I want to argue that for any dialogue about a person's mental state to transform her mental states in a non-accidental way, the person cannot know her mental state with third-personal self-knowledge. Here is a sketch of such an argument.

It is valuable to apply the heuristic I suggested earlier (IV.3.4). As I mentioned, we have clearer intuitions about how we know the mental states of a third person than we do about how we know our mental states third-personally. Because of this, when posed with a question about third-personal self-knowledge, it is helpful to first reflect on how this question is answered when we know another person's mental states and then transfer the fruits of our reflections to the case of our own third-personal knowledge.

If I have third-personal self-knowledge of Noah's belief P, then your attempts to have a dialogue with me about the merits of holding P will not make a difference to what Noah believes (unless, by accident, Noah is listening to our dialogue or I subsequently discuss what we spoke with him). The same will happen when you speak with me about beliefs about which I have third-personal self-knowledge. Your attempts to have a dialogue with me about the merits of holding P will not make a difference to what I, *qua* holder of the belief, believe. My belief will not come in direct contact with the content of our conversation. The effects that your talking with me about a mental state that I hold third-personally will have on such a mental state will be like the effects that your talking with me about Noah's mental

state will have on his mental state.

And, of course, there is no reason to think that this would be any different if the dialogue is one that I am having with myself. If I am speaking with myself about a mental state that I only know third-personally, my conversation will not come in contact with this mental state.

One might argue that this argument is not right because the proper comparison here is not to a conversation that you and I have in which Noah is absent, but one in which he is physically present, in the room with us. The objector would insist that this is the right comparison because I, *qua* holder of belief P, am physically present in the room. The objector might grant that Noah does not take part in the conversation, but insist that because he is in the room he is listening to what we say and, as a consequence, that the content of our conversation will come in direct contact with his belief. Thus, one might argue that when I have a conversation with you (or with myself) about beliefs that I know third-personally, the part of me which believes P is listening to those conversations even if I cannot speak about them *qua* speaker. And because I am there, *qua* holder of the belief, listening to these conversations, these conversations will come in contact with such a belief.

This would be a warranted objection if it was possible for me, *qua* holder of belief P, to "listen unconsciously" to this conversation. However, as I argued in chapter II, this is not possible. We can track individual words and be primed by them unconsciously, but to properly understand a conversation we need to be consciously attending to it. And because the part of me which believes P is not consciously attenting to the conversation, it is as though she is not there (or, if one wants, it is as though this part of me is there but it is mentally impaired).

Thus, when I have third-personal self-knowledge of a mental state, this mental state will be disconnected from what I can say or argue about it *qua* speaker. And this will entail that I will be impotent to change this mental state as a result of coming in contact with the arguments that I might have about them with others or with myself, i.e. with the content

of the conversation that we might have about such mental state.

Possessing first-personal self-knowledge, even if it is merely-expressive self-knowledge, changes the situation. The fact that first-personal self-knowledge allows the person's mental state to be expressed in a self-ascription entails that what the person says, thinks or hears about her self-ascription has the potential to alter the mental state that underlies this ascription. First-personal self-knowledge puts the subject in the right kind of relationship with her mental states. It enables her to transform her mental state by reflecting on the merits of holding it. To put it somewhat figuratively, first-personal self-knowledge brings the mental state into our inner conversation, it makes it possible for the mental state to listen to what the subject might have to say about it or what what the subject might learn from others about it. As I have argued in the previous paragraphs, the possibility that her mental state will listen is not available when it is known with third-personal self-knowledge.

Of course, saying that first-personal self-knowledge puts the mental state in a position to listen to the person's judgments does not mean that it will necessarily listen. But although merely-expressed self-knowledge will not, in and of itself, make a mental state effectively judgment-sensitive, it will be a step towards transforming into one, an important step to move the person from having third-personal self-knowledge to having endorsed self-knowledge.

The previous two chapters helped us see that the nature of judgment-sensitive mental states is such that, if things are going well, a judgment-sensitive mental state will be known with endorsed self-knowledge. Of course things do not always go well and there are many kind of obstacles that explain why a certain judgment-sensitive mental state will not be known with endorsed self-knowledge.

My thesis, which I plan to substantiate in future work by appealing to the clinical experience within psychotherapy, particularly to the process of working through, is that the exercise of merely-expressed self-knowledge will tend to lead to the acquisition of endorsed self-knowledge. As I said before, two important reasons why a mental state is recalcitrant,

and therefore not liable to be known with endorsed self-knowledge, is that it is emotionally difficult to hold such state or that it is cognitively expensive to do so. But if one attends to what happens within many forms of psychotherapy, these obstacles are often overcome, even if only gradually and with time. Speaking about our recalcitrant mental states with first-personal self-knowledge (i.e. exercising our capacity for merely-expressed self-knowledge) gradually leads these mental states to become effectively sensitive to our judgments. The fact that they become effectively judgment-sensitive, however, suggests that these mental states, so to speak, learn to listen to our self-conscious judgments.

The idea that mental states listen or participate in a conversation can be seen as an objectionable anthropomorphism. To the extent that this idea is meant to be more than a way of speaking, the objection is warranted. Thus it is perhaps relevant to explain how one can reformulate the issue in such a way that one does away with the anthropomorphism. A mental state is not an independent piece of mental furniture in our mind. It connects in all sort of ways with other mental states. Each individual mental state has a place in such network by the place that it plays in the life of such the organism to which the mental state can be attributed. A conflicting mental state belongs to one such network, a network that can be identified, at the very least, with a certain perspective taken by the person (perhaps unconsciously). Each of these perspectives can be identified, in turn, with the person as seen from a different angle or, if one does not find the language objectionable, with a different part of the person. My suggestion, then, is that the process of expressing, in a self-ascription, a mental state M that is conflicting will gradually lead this part of the person that is oriented by this conflicting mental state to listen to the part of the person that is oriented by her self-knowledge judgment about the lack of merits of holding M. This process is likely to be slow and perhaps to leave obdurate residues that will not be amenable to change, but it will tend to allow the person to come to hold states of mind that are more aligned with her self-conscious judgments.

VI.4.2 Am I Portraying All Therapeutic Interventions as Forms of CBT?

Readers familiar with psychotherapy might think that the process that I am describing seems to model the transformation of our mental state at play in Cognitive Behavioral Therapy (CBT). And they might complain that, in doing so, I seem to be overlooking the role that merely-expressive self-knowledge has in many other forms of psychotherapy.

This challenge is, to a certain extent, true. And below I will say more about how expressive self-knowledge is important for ethical development and how this process goes beyond the paradigmatic transformations of CBT's approach.

I want to argue, however, that while it is true that the inner dialogue that I have discussed so far is particularly salient in CBT, it is a dialogue that takes place (and ought to take place) in any other kind of therapeutic approach. There are, of course, many different activities that happen within the consulting room that are not versions of the inner dialogue that I have been describing. And it is true that many therapeutic approaches put their focus on those alternative approaches and not, like CBT, on getting our self-conscious judgment to engage and transform our recalcitrant mental states. But if it is true, as I have argued that it is, that mental health requires our mental states to be effectively judgment-sensitive, then any therapy that promotes mental health will have among its aims to make our recalcitrant mental states amenable to our self-conscious judgment. A paradigmatic way in which this happens is by a gradual transformation of our conflicting mental state that takes place as these mental states gradually listen to our self-conscious judgments.

Take the case of dynamic psychotherapy. This kind of therapy focuses on exploring and articulating mental states that are, typically, unconscious. Their being unconscious is, in fact, part of what explains their recalcitrance. As I argued in chapter V, to fully explore these kinds of mental states the person needs to bracket her deliberative perspective. Doing this

allows the irrational mental states that are informing our lives to emerge into consciousness. However, this is something that should only be done temporarily. Once the mental state has emerged into consciousness, once the person can know it with merely-expressive self-knowledge, the aim is to get the mental state to be gradually informed by the person's self-conscious judgment.

VI.4.3 The Dialogue Runs Both Ways

It is important to highlight that, within the consulting room, the dialogue between the part of the person oriented towards the conflicting mental state and the part of the person oriented by the self-conscious judgment which finds this conflicting mental state objectionable does not run in only one way.[13]

One pervasive practice within many forms of psychotherapy, often called "acceptance," involves opening ourselves to our conflicting mental states and exploring them. This exploration is not merely a third-personal exploration; it involves experiencing these conflicting mental states and seeing the world, even if briefly, through their lens.[14]

Michael Lacewing, another philosopher influenced by psychoanalysis, has suggested that this kind of exploration of our mental states, our giving them a voice, leads us to understand better these conflicting mental states, in particularly to understand better the vision of the good that they present. This often allows us to see that these mental states might not be as unmerited as we had taken them to be. Acceptance is meant, among other things, to help us realize that there are important kernels of truth in the recalcitrant mental states that we have disavowed. Thus, as Nancy Sherman points out, acceptance brings about a

13. I discussed this briefly in V.4.

14. The name "acceptance" has often lead some scholars, particularly those critical of this practice, to mischaracterize it as a literal and full-fledged acceptance of the conflicting mental state. This is a misreading. Acceptance entails a relaxation of our judgments about the merits of holding a mental state. But such relaxation is aimed at an open-minded exploration of this mental state, not to our surrendering ourselves to it by an uncritical acceptance of it.

transformation in our judgments: "judgments of what is best may not remain the same once the disavowed is reclaimed and reinterpreted"[15].

This dialogue, when properly done, has the potential to change not only our conflicting mental states but also our own judgments about their alleged merits or lack thereof. And my explanation of acceptance justifies why we should call it a dialogue and not a lecture. The process of acquiring merely-expressive self-knowledge is not merely the process whereby what I take to be my higher mental faculties inform what I take to be my lower ones, as a professor's lecture might inform the minds of her students. It is a dialogue (and not a lecture) because both parties interact and change one another, both parties are likely to be transformed by their coming into contact. Just as the recalcitrant mental state will gradually shift as it listens and is convinced by the self-conscious judgment, so will our self-conscious judgment listen and be convinced by the views put forth by our conflicting mental state, views which are warranted but which we, *qua* speakers, had been unable to recognize as such. But for this the person will need to be able to give a voice to these conflicting mental states so that this voice can engage the person's judging capacity and transform it. Merely-expressed self-knowledge provides us with such a voice.

Thus, merely-expressive self-knowledge will facilitate the process whereby conflicting mental states are controlled and gradually transformed into effectively judgment-sensitive states. An important family of such states of mind are instinctive and automatic mental states that, although implicit, shape the way in which the person perceives the world and responds to it. If a person can *only* know her mental state third-personally, she will be incapable of engaging in a conversation which has the potential to transform, in a non-accidental way, this mental state (and this will be true whether the conversation is with another person or with herself).

15. Sherman 1995, p. 233.

VI.5 Why This Transformation Is Important

I have suggested that merely-expressive self-knowledge is important in order to identify implicit patterns of thought, flesh out and articulate inchoate mental states, and develop endorsed self-knowledge of these states.

In fleshing out this last claim I argued that acquiring expressive self-knowledge was a way to develop, in a non-accidental way, endorsed self-knowledge of such state. I'd like to conclude this epilogue by briefly examining a natural challenge to this view: why is it important that we are able to transform our mental states in this way? Is there any value to the fact that the transformation is non-accidental? The objector might suggest that there might be more expedient methods to do this, if not now perhaps in the future. Perhaps we will come to a point where we have surgeons who can operate on our brain and eliminate our conflicting mental states, or perhaps psychiatrists can develop powerful drugs with no secondary effects that are able to eliminate these conflicting mental states.[16]

A natural response to this challenge is to say that by using these methods the ethical pilgrim is taking a shortcut to ethical development. And this, it might be argued, mars the merit of such development. I find this response very unsatisfactory. What the ethical pilgrim wants is to be able to live virtuously. I don't feel compelled by the thought that her development has to be meritorious. In fact, the thought that she should not appeal to these methods because they mar the merit of her progress seems to suggest that we are thinking about ethical indulgence in the wrong way; it seems to suggests a certain amount of self-indulgence in our conception of ethical development. If these methods are adequate to secure the pilgrim's ethical development, if they are more expedient and convenient, the ethical pilgrim should seek them. Her aim should be to be virtuous, not to display her moral strength through her process of ethical development.

16. The objector might actually mention that this last possibility is one that Freud hoped psychology would achieve at some point in his career (Freud 1963, p. 436)

What I intend to argue in future work is that the reason why the ethical pilgrim should not prefer to transform her mental states in an accidental way, i.e. by appealing to a brain operation or the ingestion of a certain drug, is not that these latter methods are less meritorious, but that pursuing them is a form of evasion. I plan to appeal to the clinical experience of psychotherapists and the empirical evidence uncovered by psychologists and cognitive scientists to argue that the process of transforming one's mental states in a non-accidental way involves the development and honing of virtues with which this mental state conflicts. I intend to rely on the empirical literature to argue that it is a fantasy to think that one can reach a stage of full mental harmony where one does not have conflicting mental states. Mental states are pervasive and, given how we are wired, ineliminable. Thus, I plan to argue that learning to transform these mental states is a central part of virtue. And to do so, as I have been arguing in this work, one must seek both varieties of first-personal self-knowledge: merely-expressive self-knowledge and endorsed self-knowledge.

Bibliography

Annas, Julia (1985). "Self-knowledge in early Plato". In: *Platonic Investigations*. Ed. by Dominic J. O'Meara. Washington D.C.: Catholic University of America Press, pp. 111–138.

— (1993). *The Morality of Happiness*. Oxford: Oxford University Press.

— (1995). "Virtue as a Skill". In: *International Journal of Philosophical Studies* 3.2, pp. 227–243.

— (2011). *Intelligent Virtue*. New York: Oxford University Press.

Anscombe, Elizabeth (2000). *Intention*. United States: Harvard University Press.

Antonaccio, Maria (2012). "The virtues of metaphysics: A review of Murdoch's philosohical writings". In: *Iris Murdoch, philosopher: A collection of essays*. Ed. by Justin Broackes. Oxford: Oxford University Press, pp. 155–179.

Aristotle (1999). *Metaphysics*. Translator Joe Sachs. Santa Fe, NM: Green Lion Press.

— (2002). *Nicomachean Ethics*. Trans. by Christopher Rowe. With a comment. by Sarah Broadie. New York: Oxford University Press.

Armstrong, David (1968). *A Materialist Theory of Mind*. New York: Humanities Press.

— (1981). *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Arpaly, Nomy (2000). "On Acting Rationally against One's Best Judgment". In: *Ethics* 110.3, pp. 488–513.

— (2003). *Unprincipled Virtue*. New York: Oxford University Press.

Arpaly, Nomy (2007). "Unprincipled Virtue: Synopsis (Of Sorts)". In: *Philosophical Studies* 134.3, pp. 429–431.

— (2015). "Consciousness and Moral Responsibility, by Levy, Neil". In: *Australasian Journal of Philosophy* 93.4, pp. 829–831.

Bäckström, Stina (2011). "Expression and Intentional Action". Manuscript presented in the Wittgenstein Workshop. University of Chicago. Unpublished.

Bargh, John A., ed. (2007). *The Unconscious in Social Psychology. The automaticity of higher mental processes*. New York: Psychology Press.

Bargh, John A. and Ezequiel Morsella (2008). "The Unconscious Mind". In: *Perspectives on Psychological Science* 3.1, pp. 73–79.

Barney, Rachel (2010). "Plato on the Desire for the Good". In: *Desire, Practical Reason, and the Good*. Ed. by Sergio Tenenbaum. New York: Oxford University Press, pp. 34–64.

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

Baumann, Nicola, Reiner Kaschel, and Julius Kuhl (2005). "Striving for Unwanted Goals: Stress-Dependent Discrepancies Between Explicit and Implicit Achievement Motives Reduce Subjective Well-Being and Increase Psychosomatic Symptoms". In: *Journal of Personality and Social Psychology* 89.5, pp. 781–799.

Baumeister, Roy F. and John A. Bargh (2014). "Conscious and Unconscious. Toward an Integrative Understanding of Human Mental Life and Action". In: *Dual-Process Theories of the Social Mind*. Ed. by Jeffrey W. Sherman, Bertram Gawronski, and Yaacov Trope. New York, NY: The Guilford Press. Chap. 3, pp. 35–49.

Baumeister, Roy F., E. J. Masicampo, and Kathleen D. Vohs (2011). "Do Conscious Thoughts Cause Behavior?" In: *Annual Review of Psychology* 62, pp. 331–361.

Baumeister, Roy F., Kathleen D. Vohs, and E. J. Masicampo (2014). "Maybe it helps to be conscious, after all". In: *Behavioral and Brain Sciences* 37.1, pp. 20–21.

Bechara, Antoine et al. (1997). "Deciding Advantageously Before Knowing the Advantageous Strategy". In: *Science* 275, pp. 1293–1295.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, Judith S. (2011). *Cognitive Behavioral Therapy*. New York: The Guilford Press.

Bem, Daryl J (1972). "Self-perception theory". In: *Advances in experimental social psychology* 6, pp. 1–62.

Bernacer, Javier et al. (2014). "The problem of consciousness in habitual decision making". In: *Behavioral And Brain Sciences* 37.1, pp. 21–22.

Blustein, Jeffrey (1991). *Care and Commitment: Taking the Personal Point of View*. New York: Oxford University Pres.

Bonke, B. et al. (2014). "Conscious versus unconscious thinking in the medical domain: the deliberation-without-attention effect examined". In: *Perspectives in Medical Education* 3.3, pp. 179–189.

Bos, Maarten W. and Ap Dijksterhuis (2012). "Self-Knowledge, Unconscious Thought, and Decision Making". In: *Handbook of Self-knowledge*. Ed. by Timonthy D. Wilson and Simine Vazire. New York: The Guilford Press, pp. 181–193.

Bouveresse, Jacques (1995). *Wittgenstein Reads Freud. The Myth of the unconscious*. New Jersey: Princeton University Press.

Boyle, Matthew (2009). "Two Kinds of Self-Knowledge". In: *Philosophy and Phenomenological Research* 78.1, pp. 133–164.

— (2011). "Transparent Self-Knowledge". In: *Aristotelian Society* 85 (1): *Supplementary Volume*, pp. 223–241.

Boyle, Matthew and Doug Lavin (2010). "Goodness and Desire". In: *Desire, Practical Reason, and the Good*. Ed. by Sergio Tenenbaum. New York: Oxford University Press, pp. 161–201.

Broackes, Justin (2012a). "Introduction". In: *Iris Murdoch, Philosopher*. Ed. by Justin Broackes. Oxford: Oxford University Press, pp. 1–92.

— ed. (2012b). *Iris Murdoch, Philosopher*. Oxford: Oxford University Press.

Broadie, Sarah (1987). "Nature, Craft and Phronesis in Aristotle". In: *Philosophical Topics* 15.2 (Fall), pp. 35–50.

— (1991). *Ethics with Aristotle*. New York: Oxford University Press.

Buford, Thomas O. (2011). *Know Thyself*. Lanham, Maryland: lexington Books.

Burns, Jeffrey M. and Russell H. Swerdlow (2003). "Right Orbitofrontal Tumor With Pedophilia Symptom and Constructional Apraxia Sign". In: *Arch Neurol* 60.3, pp. 437–440.

Burnyeat, Myles (1980). "Aristotle on Learning to be Good". In: *Essays on Aristotle's Ethics*. Ed. by Amelie O. Rorty. Berkley: University of California Press, pp. 69–92.

Butler, Joseph (2006). "Upon Self-Deceit". In: *The Works of Bishop Butler*. Ed. by D.E. White. Rochester: Rochester University Press.

Butler, Judith (2005). *Giving an Account of Oneself*. Fordham University Press.

Byrne, Alex (2011). "Transparency, Belief, Intention". In: *Proceedings of the Aristotelian Society* Supplementary Volume 85 (1), pp. 201–221.

Callard, Agnes (2017 (Forthcoming)). *Aspiration*. Oxford University Press.

Carruthers, Peter (2010). "Introspection: Divided and Partly Eliminated". In: *Philosophy and Phenomenological Research* 80.1, pp. 76–111.

— (2011). *The Opacity of Mind*. New York: Oxford University Press.

Cassam, Quassim (2014). *Self-Knowledge for Humans*. New York: Oxford University Press.

Castañeda, Hector-Neri (1966). "'He': A Study in the Logic of Self-Consciousness". In: *Ratio* 8, pp. 130–157.

Cooper, John (1980). "Aristotle on Friendship". In: *Essays on Aristotle's Ethics*. Ed. by Amelie O. Rorty. Berkley: University of California Press, pp. 301–340.

Cottingham, John (1998). *Philosophy and the Good Life: Reason and the Passions in Greek, Cartesian and Psychoanalytic Ethics*. New York, NY: Cambride University Press.

Cox, Danian, Marguerite La Caze, and Michel P. Levine (2003). *Integrity and the fragile self*. Burlington, USA: Ashgate.

Crockett, Molly J. et al. (2008). "Serotonin modulates behavioral reactions to unfairness." In: *Science* 320.5884, p. 1739.

D'Arms, Justin and Daniel Jacobson (2000a). "Sentiment and Value". In: *Ethics* 110.4, pp. 722–748.

— (2000b). "The Moralistic Fallacy: On the 'Appropriateness' of Emotions". In: *Philosophy and Phenomenological Research* 61.1, pp. 65–90.

Davidson, Donald (1982). "Paradoxes of Irrationality". In: *Philosophical Essays on Freud*. Ed. by Richard Wollheim and J. Hopkins. Cambridge: Cambridge University Press.

Deutsch, Roland, Bertram Gawronski, and Fritz Strack (2006). "At the Boundaries of Automaticity: Negation as Reflective Operation". In: *Journal of Personality and Social Psychology* 91.3, pp. 385–405.

DeWall, Nathan, Roy F. Baumeister, and E. J. Masicampo (2008). "Evidence that logical reasoning depends on conscious processing". In: *Consciousness and Cognition* 17.3, pp. 628–645.

Deweese-Boyd, Ian (2012). "Self-Deception". In: *The Stanford Encyclopedia of Philosophy* (Spring). Ed. by Edward N. Zalta. URL: `http://plato.stanford.edu/archives/spr2012/entries/self-deception/`.

Dijksterhuis, Ap, Maarten W. Bos, et al. (2006). "On making the right choice: The deliberation-without-attention effect." In: *Science* 311.5763, pp. 1005–1007.

Dijksterhuis, Ap and Loran F. Nordgren (2006). "A Theory of Unconscious Thought". In: *Perspectives on Psychological Science* 1.2, pp. 95–109.

Dobson, Keith and Renee-Louise Franche (1989). "A conceptual and empirical review of the depressive realism hypothesis". In: *Canadian Journal of Behavioral Sciences* 21.4, pp. 419–433.

Doris, John (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambride University Press.

Doris, John M. (2009). "Skepticism About Persons". In: *Philosophical Issues* 19.1, pp. 57–91.

Dougherty, Kathleen Poorman (2000). "Self-knowledge and Moral Virtue". PhD. Thesis. University of Oklahoma.

Driver, Julia (1989). "The Virtues of Ignorance". In: *The Journal of Philosophy* 86.7, pp. 373–394.

— (1999). "Modesty and Ignorance". In: *Ethics* 109.4, pp. 827–834.

— (2001). *Uneasy Virtue*. New York, NY: Cambridge University Press.

Edgley, Roy (1969). *Reason in theory and practice*. London: Hutchinson.

Engstrom, Stephen (2007). "Kant on the Agreeable and the Good". In: *Moral psychology*. Ed. by Sergio Tenenbaum. New York: Rodopi, pp. 111–160.

— (2009). *The Form of Practical Knowledge*. Cambridge, Massachusetts: Harvard University Press.

Estabrooks, George. H. (1943). *Hypnotism*. New York: E. P. Dutton & Company.

Evans, Gareth (1982). *The Varieties of Reference*. Ed. by John McDowell. Oxford: Oxford University Press.

Evans, Jonathan St. B. T. (2008). "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition". In: *Annual Review of Psychology* 59.

Falkenström, Fredrik (2012). *The Capacity for Self - Observation in Psychotherapy*. Linköping, Sweeden: Department of Behavioural Sciences and Learning Linköping University.

Finkelstein, David (1999). "On The Distinction Between Conscious And Unconscious States of Mind". In: *American Philosophical Quarterly* 36.2, pp. 79–100.

Finkelstein, David (2000). "Wittgenstein on Rules and Platonism". In: *The New Wittgenstein*. Ed. by Alice Crary and Rupert Read. New York: Routledge.

— (2001). "Wittgenstein's 'Plan for the treatment of psychological concepts'". In: *Wittgenstein in America*. Ed. by Timothy McCarthy and Sean Stidd. Oxford, pp. 215–236.

— (2003). *Expression and the Inner*. Cambridge, Massachusetts: Harvard University Press.

Flanagan, Owen (1990). "Virtue and Ignorance". In: *The Journal of Philosophy* 87.8, pp. 420–428.

— (1991). *Varieties of Moral Personality*. Cambridge, MA: Harvard University Press.

Floyd, Juliet (1995). "On Saying what you really want to say: Wittgenstein, Gödel, and the Trisection of the Angle". In: *From Dedekind to Gödel. Essays on the development of the foundations of mathematics*. Ed. by Kaakko Hintikka. Norwell, Massachusetts: Kluwer Academic Publishers, pp. 373–426.

— (2000). "Wittgenstein, mathematics and philosophy". In: *The New Wittgenstein*. Ed. by Crary Alice and Rupert Read. New York: Routledge, pp. 232–261.

Ford, Anton (2011). "Action and generality". In: *Essays on Anscombe's Intention*. Ed. by Jennifer Hornsby Anton Ford and Frederick Stoutland. Cambridge, Massachusetts: Harvard University Press.

Foucault, Michel (1988). *Technologies of the Self*. Ed. by Luther H. Martin. Amherst: The University of Massachusetts Press.

— (2005). *The Hermeneutics of the Subject*. New York: Picador.

Frankfurt, Harry G. (1988). *The Importance of What We Care About*. Cambridge University Press.

Frankish, Keith and Jonathan St. B. T. Evans (2009). "The duality of mind: an historical perspective". In: *In Two Minds: Dual Processes and Beyond*. Ed. by Keith Frankish and Jonathan St B. T. Evans. New York: Oxford University Press.

Freud, Sigmund (1925). “Negation”. In: *The International Journal of Psychoanalysis* 6, pp. 367–371.

— (1953a). “Notes Upon a Case of Obsessional Neurosis”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 10. Great Britain: The Hogarth Press.

— (1953b). “The Unconscious”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 14. Great Britain: The Hogarth Press.

— (1955a). “Studies on Hysteria”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 16. Great Britain: The Hogarth Press.

— (1955b). “The ‘Uncanny’”. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Ed. by James Strachey. Vol. 17. Great Britain: The Hogarth Press, pp. 217–256.

— (1958a). “A Note on the Unconscious in Psychoanalysis”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 255–266.

— (1958b). “Observations on transference love (Further recommendations on the technique of psychoanalysis III)”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 157–171.

— (1958c). “On Beginning the Treatment”. In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 121–144.

Freud, Sigmund (1958d). "Recommendations to Physicians Practising Psycho-Analysis". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 109–120.

— (1958e). "The Dynamics of Transference". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 97–108.

— (1958f). "Two Principles of Mental Functioning". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 215–226.

— (1962). "On the history of the Psycho-analytic Movement". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 14. Great Britain: The Hogarth Press, pp. 3–66.

— (1963). "Introductory Lectures on Psychoanalysis. (Part III.)" In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Ed. by James Strachey. Vol. 16. Great Britain: The Hogarth Press.

— (1964). "An Outline of psycho-analysis". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 23. Great Britain: The Hogarth Press.

— (1980). "Remembering, Repeating and Working-Through". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 12. Great Britain: The Hogarth Press, pp. 145–156.

— (1995a). "The Future Prospects of Psycho-analytic Therapy". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 11. Great Britain: The Hogarth Press, pp. 139–152.

Freud, Sigmund (1995b). "Two Encyclopaedia Articles". In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Ed. by James Strachey. Vol. 18. Great Britain: The Hogarth Press, pp. 233–254.

— (1995c). "Wild Psycho-Analysis". In: *The Standard Edition of the Complete psychological works of Sigmund Freud*. Ed. and trans. by James Strachey. Vol. 11. Great Britain: The Hogarth Press, pp. 219–230.

Friese, Malte, Wilhelm Hofmann, and Michaela Wänke (2008). "When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour". In: *British Journal of Social Psychology* 47, pp. 397–419.

Gabbard, Glen O. and Drew Westen (2003). "Rethinking Therapeutic Action". In: *International Journal of Psycho-Analysis,* 84.4, pp. 823–841.

Gardner, Sebastian (1993). *Irrationality and the Philosophy of Psychoanalysis*. New York: Cambridge University Press.

Gawronski, Bertram and Galen V. Bodenhausen (2012). "Self-Insight from a Dual-Process Perspective". In: *Handbook of Self-knowledge*. Ed. by Timonthy D. Wilson and Simine Vazire. New York: The Guilford Press, pp. 22–38.

Gawronski, Bertram and Laura A. Creighton (2013). "Dual Process Theories". In: *The Oxford Handbook of Social Cognition*. Ed. by D. E. Carlston. New York: Oxford University Press. Chap. 14.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J. Wilbur (2006). "Are 'Implicit' Attitudes Unconscious?" In: *Consciousness and Cognition* 15, pp. 485–99.

Gawronski, Bertram and Fritz Strack (2004). "On the Propositional Nature of Cognitive Consistency: Dissonance Changes Explicit, But Not Implicit Attitudes". In: *Journal of Experimental Social Psychology* 40.4, pp. 535–42.

Gazzaniga, Michael (1995). "Consciousness and the cerebral hemispheres". In: *The Cognitive Neurosciences*. Ed. by Michael Gazzaniga. MIT Press.

— (2000). "Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition?" In: *Brain* 123, pp. 1293–1326.

Gendler, Tamar S. (2008a). "Alief And Belief". In: *The Journal of Philosophy* 105 (10), pp. 634–663.

— (2008b). "Alief in Action (and Reaction)". In: *Mind & Language* 23 (5), pp. 552–585.

— (2014). "The Third Horse: On Unendorsed Association and Human Behaviour". In: *Proceedings of the Aristotelian Society* Supplementary Volume 88 (1), pp. 185–218.

Gerson, Lloyd P. (1994). *Plotinus*. New York: Routledge.

— ed. (1996). *Cambridge Companion to Plotinus*. 1st ed. Australia: Cambridge University Press.

Gertler, Brie (2011). "Self-Knowledge". In: *The Stanford Encyclopedia of Philosophy* (Spring). Ed. by Edward N. Zalta. URL: `http://plato.stanford.edu/archives/spr2011/entries/self-knowledge/`.

gertler, Brie (2011). *Self-Knowledge*. New York: Routledge.

Gill, Christopher (2007). "Self-knowledge in Plato's Alcibiades". In: *Reading Ancient Texts, Vol 1: Presocratics and Plato, Essays in honour of Denis O'Brien*. Ed. by S. Stern-Gillet and K. Corrigan. Leiden: Brill, pp. 97–112.

Girenzer, Greg (2007). *Gut Feelings. The Intelligence of the Emotions*. New York: Viking.

Gladwell, Malcom (2005). *Blink, the power of thinking without thinking*. Little Brown.

Goldie, Peter (2000). *The Emotions: A Philosophical Exploration*. Oxford: Oxford University Press.

— (2009). "Narrative Thinking, Emotion, and Planning". In: *Journal of Aesthetics and Art Criticism* 1.97, pp. 97–106.

— (2010). *The Oxford Handbook of Philosophy of Emotion*. Oxford: Clarendon Press.

Gopnik, Alison (1993). "How we know our minds: The illusion of first-person knowledge of intentionality". In: *Behavioral and Brain Sciences* 16 (1), pp. 1–14.

Gordon, Peter C., Randall Hendrick, and William H. Levine (2002). "Memory-Load Interference In Syntactic Processing". In: *Psychological Science* 13.5, pp. 425–430.

Gray, Paul (1982). ""Developmental Lag" in the Evolution of Technique for Psychoanalysis of Neurotic Conflict". In: *Journal of the American Psychoanalytic Association* 30, pp. 625–655.

Green, Mitchell S. (2008). *Self-expression*. New York: Oxford University Press.

Grenberg, Jeanine (2005). *Kant and the ethics of humility*. New York: Cambridge University Press.

Gurman, Alan S. and Stanley B. Messer, eds. (2003). *Essential Psychotherapies*. New York: The Guilford Press.

Haase, Matthias. "Practically Self-Conscious Life". (Unpublished Manuscript).

Hadot, Pierre (1993). *Plotinus or the Simplicity of Vision*. Chicago: The University of Chicago Press.

— (1995). *Philosophy as a Way of Life*. Ed. by Arnold I. Davidson. Oxford: Blackwell.

— (1998). *The Inner Citadel*. Trans. by Michael Chase. London: Harvard University Press.

Hampshire, Stuart (1982). "Disposition and Memory". In: *Philosophical essays on Freud*. Ed. by Richard Wollheim and James Hopkins. New York: Cambridge University Press, pp. 75–91.

Hansen, Katherine E. and Emily Pronin (2012). "Illusions of Self-knowledge". In: *The Handbook of Self-knowledge*. Ed. by Timothy D. Wilson and Simine Vazire. New York: The Guilford Press. Chap. 21, pp. 345–362.

Hassin, Ran R. and Maxim Milyavsky (2014). "But what if the default is defaulting". In: *Behavioral and Brain Sciences* 37.1, pp. 29–30.

Hassin, Ran R., James S. Uleman, and John A. Bargh (2005). *The New Unconscious*. New York: Oxford University Press.

Hayes, Steven C., Kirk D Strosahl, and Kelly G. Wilson (2003). *Acceptance and Commitment Therapy*. New York: The Guilford Press.

Herman, Barbara (1993). *The Practice of Moral Judgement*. United States of America: Harvard University Press.

Hills, Allison (2015). "The Intellectuals and the Virtues". In: *Ethics* 126.1, pp. 7–36.

Hofer, Jan, Athanasios Chasiotis, and Domingo Campos (2006). "Congruence Between Social Values and Implicit Motives: Effects on Life Satisfaction Across Three Cultures". In: *European Journal of Personality* 20, pp. 305–324.

Holland, Margeret (2012). "Social Convention and Neurosis as obstacles to Moral freedom". In: *Iris Murdoch, Philosopher*. Ed. by Justin Broackes. Oxford: Oxford University Press, pp. 255–273.

Hume, David (1998). *An Enquiry Concerning the Principles of Morals*. Ed. by Tom L. Beauchamp. Oxford: Oxford University Press.

— (2007). *A Treatise of Human Nature: A Critical Edition*. Ed. by David Fate Norton and Mary J. Norton. Oxford: Clarendon Press.

Hursthouse, Rosalind (1991). "Arational Actions". In: *The Journal of Philosophy* 88.2, pp. 57–68.

— (2002). *On Virtue Ethics*. New York: Oxford University Press.

Ingram, Douglas H (1982). "Compulsive Personality Disorder". In: *American Journal of Psychoanalysis* 42, pp. 189–198.

Jacobs, Jonathan (1989). *Virtue and Self-Knowledge*. Prentice Hall.

James, William (1950a). *The Principles of Psychology, Vol. 1*. Dover Publications.

— (1950b). *The Principles of Psychology, Vol. 1*. Dover Publications.

James, William (1983). "The Hidden Self". In: *Essays in Psychology*. Ed. by Frederick H Burkhardt, Fredson Bowers, and Ignas Skrupkelis. Cambridge, Mass: Harvard University Press, pp. 247–265.

— (2002). *The Varieties of Religious Experience*. Dover Publications.

Jeannerod, Marc (2006). "Consciousness of action as an embodied consciousness". In: *Does Consciousness Cause Behavi*. Ed. by William P. Banks Susan Pockett and Shaun Gallagher. Cambridge, MA: MIT Press.

Johnson, David M. (1999). "God as the True Self: Plato's *Alcibiades*". In: *Ancient Philosophy* 19, pp. 1–19.

Jolley, Kelly Dean (2014). "Disposable Thinking". Paper presented in the Literature and Philosophy workshop, University of Chicago.

Jopling, David (2000). *Self-knowledge and the Self*. New York: Routledge.

— (2008). *Talking Cures and Placebo Effects*. International Perspectives in Philosophy and Psychiatry. New York: Oxford University Press.

Kant, Immanuel (1996). *The Metaphysics of Morals*. Translator Mary J. Gregor. Cambridge: Cambridge University Press.

Kaufmann, Walter (1950). "Nietzsche: Philosopher, Psychologist, Antichrist". In: Binghamton, N.Y.: Princeton University Press. Chap. 1.

Kellog, Scott H. and Jeffrey E. Young (2008). "Cognitive Therapy". In: *Twenty-First Century Psychotherapies*. Ed. by Jay L. Lebow. Hoboken, NJ: John Wiley & Sons. Chap. 3.

Korsgaard, Christine M. (1989). "Personal Identity and the Unity of Agency: A Kantian Response to Parfit". In: *Philosophy and Public Affairs* 18.2, pp. 101–132.

— (1999). "Self-constitution in the ethics of Plato and Kant". In: *The journal of ethics* 3, pp. 1–29.

— (2000). *The Sources of Normativity*. Ed. by Onora O'neill. Cambridge, U. K.: Cambridge University Press.

Korsgaard, Christine M. (2009). *Self–Constitution, Agency, Identity, and Integrity*. Oxford university press.

Lacewing, Michael (2008). "What Reason Can't Do". In: *The moral life: essays in honour of John Cottingham*. Ed. by Nafsika Athanassoulis and Samantha Vice. Palgrave Macmillan, pp. 139–163.

— (2013). "Psychoanalysis, Emotions and Living a Good Life". In: *Think* 12.33 (Spring), pp. 41–51.

— (2014). "Emotions and the Virtues of Self-Understanding". In: *Emotion and Value*. Ed. by Sabine Roeser and Cain Todd. Oxford University Press, pp. 199–211.

Laplanche, Jean (1985). *Life and Death in Psychoanalysis*. Johns Hopkins University Press.

Laplanche, Jean and J. B. Pontalis (1973). *The Language of Psycho-analysis*. W. W. Norton & Company.

Larigauderie, Pascale, Daniel Gaonac'h, and Natasha Lacroix (1998). "Working memory and error detection in texts: what are the roles of the central executive and the phonological loop?" In: *Applied Cognitive Psychology* 12.5, pp. 505–527.

Laurence, Ben (2012). "The Priority of Ideal Theory". Presented at the Practical Philosophy Workshop.

Lavin, Douglas. (2004). "Practical Reason and the Possibility of Error". In: *Ethics* 114.3, pp. 424–457.

Lear, Jonathan (1990). *Love and Its Place in Nature*. New Haven: Yale University Press.

— (1998). *Open Minded: working out the logic of the soul*. Cambridge, Massachusetts: Harvard University Press.

— (2000). *Happiness, Death and the Remainder of Life*. Cambridge, Massachusetts: Harvard University Press.

— (2003). *Therapeutic Action. An Earnest Plea for Irony*. New York: Other Press.

— (2006). *Freud*. New York: Routledge.

Lear, Jonathan (2012). "A Lost Conception of Irony". In: *Berfrois*.

Lear, Jonathan et al. (2011). *A Case for Irony*. Cambridge, Massachusetts: Harvard University Press.

Leiter, Brian (2007). "Nietzsche's Theory of the Will". In: *Philosopher's Imprint* 7.7, pp. 1–15.

Levy, Neil (2013). "The Importance of Awareness". In: *Australasian Journal of Philosophy* Vol. 91.2, pp. 211–229.

— (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Lieberman, Matthew D. (2012). "Self-Knowledge: From Philosophy to Neuroscience to Psychology". In: *Handbook of Self-knowledge*. Ed. by Timonthy D. Wilson and Simine Vazire. New York: The Guilford Press, pp. 63–76.

Livingstone Smith, David (1999). *Freud's philosophy of the unconscious*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Loewald, Hans (1980). "On Therapeutic Action in psychoanalysis". In: *Papers on Psychoanalysis*. New Haven: Yale University Press, ???

Long, A. A., ed. (1996). *Stoic Studies*. Cambridge: Cambridge University Press.

Lycan, William D. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

MacIntyre, Alasdair (2002). *Dependent Rational Animals*. Chicago: Open Court.

— (2007). *After Virtue*. 3rd ed. Notre Dame: The University of Notre Dame Press.

Martin, Mike W. (1986). *Self-deception and Morality*. Lawrence, Kansas: University Press of Kansas.

— (2006). *from morality to mental health*. New York: Oxford University Press.

McDowell, John (1998a). *Mind, Value, and Reality*. Cambridge, Massachusetts: Harvard University Press.

— (1998b). "Reductionism and the First Person". In: *Mind, Value, and Reality*. Cambridge, Massachusetts: Harvard University Press.

McDowell, John (1998c). "Virtue and Reason". In: *Mind, Value, and Reality*. Cambridge, Massachusetts: Harvard University Press.

McGeer, Victoria (1996). "Is "Self-Knowledge" An Empirical Problem? Renegotiating the Space of Philosophical Explanation". In: *The Journal of Philosophy* 93.10, pp. 483–515.

— (2007). "The Moral Development of First-Person Authority". In: *European Journal of Philosophy* 16.1, pp. 81–108.

McGeer, Victoria and Philip Pettit (2002). "The Self-Regulating Mind". In: *Language & Communication* 22 (3), pp. 281–299.

McGinn, Collin (1982). *The Character of Mind*. Oxford University Press.

McKinnon, Christine (1991). "Hypocrisy, With a Note on Integrity". In: *American Philosophical Quarterly* 28.4, pp. 321–30.

Meissner, W. W. (2003). *The Ethical Dimension of Psychoanalysis*. Albany: State of New York Press.

Merritt, Maria (2000). "Virtue ethics and situationist personality psychology". In: *Journal of Moral Education* 3, pp. 365–383.

Millgram, Elijah. "Who Wrote Nietzsche's Autobiography?" (Unpublished Manuscript).

Mitchell, Stephen A. and Margeret J. Black (1995). *Freud and Beyond. A History of modern psychoanalytic thought*. New York: Basic Books.

Mole, Christopher (2007). "Attention, Self and The Sovereignty of Good". In: *Iris Murdoch. A Reassessment*. Ed. by Anne Rowe. New York: Palgrave Macmillan, pp. 72–84.

Moore, Michael T. and David M. Fresco (2012). "Depressive realism: A meta-analytic review". In: *Clinical Psychology Review* 32, pp. 496–509.

Moran, Richard (2001). *Authority and Estrangement*. New Jersey: Princeton University Press.

Moran, Richard. (2002). "Frankfurt on identification: Ambiguities of activity on mental life." In: *The Contours of Agency: Essays on Themes from Harry Frankfurt*. Ed. by Sarah Buss and Lee Overton. The MIT Press.

Moran, Richard (2011). "Psychoanalysis and the limits of reflection". In: Lear, Jonathan. *A Case for Irony*. Cambridge, Massachusetts: Harvard University Press. Chap. 5, pp. 103–114.

— (2012). "Iris Murdoch and Existentialism". In: *Iris Murdoch, Philosopher: A Collectoin of Essays*. Ed. by Justin Broackes. New York: Oxford University Press, pp. 181–196.

Murdoch, Iris (1993). *Metaphysics as a Guide to Morals*. New York, NY: Penguin Books.

— (1997a). *Existentialists and Mystics*. Ed. by Peter Conradi. New York: Penguin group.

— (1997b). "On 'God' and 'Good'". In: *Existentialists and Mystics*. Ed. by Peter Conradi. New York: Penguin group, pp. 337–362.

— (1997c). "The Idea of Perfection". In: *Existentialists and Mystics*. Ed. by Peter Conradi. New York: Penguin group, pp. 299–336.

— (1997d). "The Sovereignity of the Good Over Other Concepts". In: *Existentialists and Mystics*. Ed. by Peter Conradi. New York: Penguin group, pp. 363–385.

— (2000). *The Nice and the Good*. Kindle Edition. London: Vintage.

— (2001). *The Bell*. Kindle Edition. New York: Penguin Classics.

Murdoch, Iris et al. (1963). "Freedom and knowledge". In: *Freedom and the will*. Ed. by David F. Pears. New York: St. Martin's Press, pp. 80–104.

Nagel, Jennifer (2014). "Intuition, Reflection, and The Command of Knowledge". In: *Proceedings of the Aristotelian Society* Supplementary Volume 88 (1), pp. 219–241.

Nagel, Thomas (1971). "Brain Bisection and the Unity of Consciousness". In: *Synthese* 22.3–4, pp. 396–413.

Narvaez, Darcia and Kellen Mrkva (2014). "The Development of Moral Imagination". In: *The Ethics of Creativity*. Ed. by Seana Moran, David Cropley, and James C. Kaufman. New York, NY: Palgrave McMillan. Chap. 1, pp. 25–45.

Newell, Ben R. and David R. Shanks (2014). "Unconscious influences on decision making: A critical review". In: *Behavioral and Brain Sciences* 37, pp. 1–19, 45–53.

Neys, Wim De (2006). "Two Systems but One Reasoner. Dual Processing in Reasoning". In: *Psychological Science* 17.5, pp. 428–433.

Nietzsche, Friedrich (1974). *The Gay Science*. Translator Walter Kaufmann. New York: Random House.

— (1998). *On the Genealogy of Morality*. Trans. by Maudemarie Clark and Alan J. Swensen. Indianapolis, IN: Hackett Publishing Company.

Nieuwenstein, M. R. et al. (2015). "On making the right choice: a meta-analysis and large-scale replication attempt of the unconscious thought advantage." In: *Judgment and Decision Making* 10, pp. 1–17.

Nisbett, Richard E. and Timothy D. Wilson (1977). "Telling More Than We Can Know: Verbal Reports On Mental Processes". In: *Psychological Review* 84.3, pp. 231–259.

Nussbaum, Martha C. (1988). "Non-Relative Virtues: An Aristotelian Approach". In: *Midwest studies in philosophy* 13.1, pp. 32–53.

— (1990a). "Love's Knowledge". In: *Love's Knowledge*. New York: Oxford University Press. Chap. 11.

— (1990b). *Love's Knowledge*. New York: Oxford University Press.

— (1994). *The Therapy of Desire*. Princeton, New Jersey: Princeton University Press.

— (1999). "Virtue Ethics: A Misleading Category?" In: *Ethics* 3, pp. 163–201.

— (2001). *Upheavals of Thought*. New York: Cambridge University Press.

— (2006). *Hiding from Humanity: Disgust, Shame, and the Law*. Princeton, New Jersey: Princeton University Press.

O'Brien, Lucy (2003). "Moran on Agency and Self-Knowledge". In: *European Journal of Philosophy* 11, pp. 375–390.

— (2005). "Self-Knowledge, Agency and Force". In: *Philosophy and Phenomenological Research* 71.3, pp. 580–601.

Oliner, Samuel P. and Pearl M. Oliner (1988). *The Altruistic Personality*. New York: The Free Press.

O'neill, Onora (1998). "Kant's Virtues". In: *How Should One Live? Essays on the Virtues*. New York: Oxford University Press, pp. 77–97.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Clarendon Press.

— (2003). "Why our identity is not what matters". In: *Personal Identity*. Ed. by Raymond Martin and John Barresi. Malden, USA: Blackwell Publishing, pp. 115–143.

Parkes, Graham (1989). "A Cast Of Many: Nietzsche And Depth-Psychological Pluralism." In: *Man and World: An International Philosophical Review* 22, pp. 453–470.

PDM Task Force (2006). *Psychodynamic Diagnostic Manual*. Silver Spring, Maryland: Alliance of Psychoanalytic Organizations.

Pears, David (1984). *Motivated Irrationality*. New York: Oxford University Press.

Pennebaker, James. W. and Cindy K. Chung (2014). "Expressive writing and its links to mental and physical health". In: *Oxford handbook of health psychology*. Ed. by Howard S. Friedman. New York: Oxford University Press, pp. 417–437.

Pippin, Robert (2001). "Morality as Psychology; Psychology as Morality: Nietzsche, Eros and Clumsy Lovers". In: *Nietzsche's Postmoralism: Essays on Nietzsche's Prelude to Philosophy's Future*. Ed. by Richard Schacht. Cambridge: Cambridge University Press, pp. 79–99.

— (2007). "Can There Be 'Unprincipled Virtue'? Comments on Nomy Arpaly". In: *Philosophical Explorations* 10.3, pp. 291–301.

Plato (1987). *Gorgias*. Translator Donald Zeyl. Indianapolis, IN: Hackett Publishing Company.

Pollard, Bill (2003). "Can virtuous actions be both habitual and rational?" In: *Ethical Theory Moral Practice* 6.4, pp. 411–425.

Pronin, Emily (2009). "The Introspection Illusion". In: *Advances in Experimental Social Psychology* 41, pp. 1–66.

Radford, Colin (1978). "It's on the tip of my tongue". In: *Philosophical Investigations* 1.2, pp. 70–79.

Rappe, Sara (1996). "Self-Knowledge and subjectivity in the *Enneads*". In: *Cambridge Companion to Plotinus*. Ed. by Lloyd P. Gerson. New York: Cambridge University Press, pp. 250–274.

Raterman, Ty (2006). "On Modesty: Being Good and Knowing It without Flaunting It". In: *American Philosophical Quarterly* 43.3, pp. 221–234.

Reinecke, Mark A. and Arthur Freeman (2003). "Cognitive Therapy". In: *Essential Psychotherapies*. Ed. by Alan S. Gurman and Stanley B. Messer. The Guilford Press. Chap. 7.

Ricoeur, Paul (1973). "The Model of the Text: Meaningful Action Considered as a Text". In: *New Literary History* 5.1, pp. 91–117.

— (1995). *Oneself as Another*. Trans. by Kathleen Blamey. Chicago: The University of Chicago Press.

Rider, Benjamin A. (2011). "Self-Care, Self-Knowledge, and Politics in the Alcibiades I". In: *Epoché* 15.2, pp. 395–413.

Rödl, Sebastian (2003). "Three forms of Practical Reasoning". (Unpublished Manuscript).

— (2007). *Self-Consciousness*. Cambridge, Massachusetts: Harvard University Press.

Rorty, Amélie Oksenberg (1994). "User-Friendly Self-Deception". In: *Philosophy* 69.268, pp. 211–228.

Rorty, Amélie Oksenberg and Owen Flanagan, eds. (1993). *Identity, Character, and Morality: Essays in Moral Psychology*. Cambridge: MIT press.

Rosenthal, David M. (2008). "Consciousness and its function". In: *Neuropsychologia* 46, pp. 829–840.

Ryle, Gilbert (1949). *The Concept of Mind*. New York: Barnes and Noble.

— (2009). "Teaching and training". In: *Collected Essays 1929–1968*. Vol. 2. New York: Routledge, pp. 464–478.

Samuels, Steven and William Casebeer (2005). "A social psychological view of morality: why knowledge of situational influences on behavior can improve character development practices". In: *Journal of Moral Education* 34.1, pp. 73–87.

Sarkissian, Hagop (2010). "The Problems and Promise of Situationism in Moral Philosophy". In: *Philosophers' Imprint* 10.9, pp. 1–15.

Scanlon, Thomas M. (2002). "Reasons and Passions". In: *Contours of Agency: Essays on Themes from Harry Frankfurt*. Ed. by Sarah Buss and Lee Overton. Cambridge, Ma: MIT Press, pp. 165–183.

Schoeller, Donata (2013). "Emerging meanings and values. Dialectic, Pragmatism and Psychotherapy". Presented at the Social Thought Colloquium on February 05, 2013.

Schultheiss, Oliver C. and Alexandra Strasser (2012). "Referential Processing and Competence as Determinants of Congruence between implicit and explicit motives". In: *Handbook of Self-knowledge*. Ed. by Timonthy D. Wilson and Simine Vazire. New York: The Guilford Press, pp. 39–62.

Segal, Hana (1974). *Introduction to the work of Melanie Klein*. New York: Basic Books Inc.

Shah, Nishi and J. David Velleman (2005). "Doxastic Deliberation". In: *The Philosophical Review* 114.4, pp. 497–534.

Shapiro, David (1965). *Neurotic Styles*. New York: Basic Books, Inc.

Shapiro, David (2001). "OCD or Obsessive-Compulsive Character?" In: *Psychoanalytic Inquiry* 21, pp. 242–252.

Sherman, Nancy (1987). "Aristotle on Friendship and the Shared Life". In: *Philosophy and Phenomenological Research* 47.4, pp. 589–613.

— (1991). *The Fabric of Character*. New York: Oxford Univeristy Press.

— (1993). "Wise Maxims/Wise Judging". In: *The Monist* 76.1, pp. 41–65.

— (1995). "The Moral Perspective and The Psychoanalytic Quest". In: *Journal of American Academy of Psychoanalysis* 23, pp. 223–241.

— ed. (1998). *Aristotle's Ethics: critical essays*. Maryland: Rowman and Littlefield Publishers.

— (2007). "Virtue and a Warrior's Anger". In: *Working virtue : virtue ethics and contemporary moral problems*. Ed. by Rebecca L. Walker and Philip J. Ivanhoe. Oxford University Press, pp. 251–277.

— (2014a). "Hope After War". Paper presented in the Practical Philosophy Workshop, University of Chicago.

— (2014b). "Recovering lost goodness: Shame, guilt, and self-empathy." In: *Psychoanalytic Psychology* 31.2, pp. 217–235.

Silbermann, I. (1985). "On 'Happiness'". In: *The Psychoanalityc Study of the Child* 40, pp. 457–472.

Small, Will (2012). "Transmission of Skill". Unpublished Manuscript.

Smith, Angela M. (2005). "Responsibility for Attitudes: Activity and Passivity in Mental Life". In: *Ethics* 115.2, pp. 236–271.

— (2008). "Control, Responsibility, and Moral Assessment". In: *Philosophical Studies* 138.3, pp. 367–392.

Smith, Eliot R. and Frederick D. Miller (1978). "Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson". In: *Psychological Review* 4.85, pp. 355–362.

Snow, Nancy (2016). "From 'Ordinary' Virtue to Aristotelian Virtue". Unpublished manuscript presented at the 4 th Annual Jubilee Centre for Character and Virtues conference. Oriel College, Oxford University.

Snow, Nancy E. (2006). "Habitual Virtuous Actions and Automaticity". In: *Ethical Theory and Moral Practice* 9.5, pp. 545–561.

Solomon, Robert (1999). "The Philosophy of Emotions". In: *Handbook of Emotions*. Ed. by Mark Lewis and Jeannette Haviland-Jones. New York: Guilford Press, pp. 3–15.

— (2003). *What is an Emotion? Classic and Contemporary Readings*. New York: Oxford University Press.

— (2004). *In Defense of Sentimentality*. New York: Oxford University Press.

Sorabji, Richard (1980). "Aristotle on the Role of Intellect in Virtue". In: *Essays on Aristotle's Ethics*. Ed. by Amelie O. Rorty. Berkley: University of California Press, pp. 201–220.

— (2006). *Self. Ancient and Modern Insights about Individuality, Life and Death*. Chicago: The University of Chicago Press.

Sousa, Ronald de (1987). *The Rationality of Emotion*. Cambridge, MA: MIT press.

Sousa, Ronald de and Adam Morton (2002). "Emotional Truth". In: *Proceedings of the Aristotelian Society* 76: *Supplementary Volumes*, pp. 247–263, 265–275.

Springsted, Eric O. (2004). "I Dreamed I saw St. Augustine . . . " In: *The Christian Platonism of Simone Weil*. Ed. by E. Jane Doering and Eric O. Springsted. Notre Dame, Indiana: University of Notre Dame press, pp. 209–228.

Stohr, Karen E. (2003). "Moral Cacophony: When Continence is a Virtue". In: *The Journal of Ethics* 7 (4), pp. 339–363.

Strube, Michael J. (2012). "From 'Out There' to 'In Here'. Implications of Self-Evaluation Motives for Self-Knowledge". In: *Handbook of Self-knowledge*. Ed. by Timonthy D. Wilson and Simine Vazire. New York: The Guilford Press, pp. 397–412.

Taylor, Gabriele (1981). "Integrity". In: *Proceedings of the Aristotelian Society* 55, pp. 143–59.

— (1985). *Pride, Shame and Guilt: Emotions of Self-assesment*. Oxford: Oxford University Press.

Thomas, Alan (2005). "Reasonable Partiality and the agent's point of view". In: *Ethical Theory and Moral Practice* 8, pp. 25–43.

Thompson, Michael (2004a). "Apprehending Human Form". In: *Modern moral philosophy*. Ed. by Anthony O'Hear. New York: Cambridge University Press.

— (2004b). "What is it to wrong someone: a puzzle about justice". In: *Reason and Value: Themes from the Moral Philosophy of Joseph Raz. Clarendon Press*. Ed. by R. Jay Wallace et al. Clarendon Press, pp. 333–384.

— (2008a). *Life and Action : Elementary Structures of Practice and Practical Thought*. Harvard, Massachusetts: Harvard University Press.

— (2008b). "The Representation of life". In: *Life and Action: Elementary Structures of Practice and Practical Thought*. Harvard, Massachusetts: Harvard University Press.

Trianosky, Gregory W. (1988). "Rightly Ordered Appetites: How to Live Morally and Live Well". In: *American Philosophical Quarterly* 25.1, pp. 1–12.

Vadillo, M. A., O. Kostopoulou, and D. R. Shanks (2015). "A critical review and meta-analysis of the unconscious thought effect in medical decision making". In: *Frontiers in Psychology* 6.636.

Valentine, Elizabeth R. (1982). "Conceptual Issues in Psychology". In: Winchester, Massachusets: George Allen & Unwin Ltd. Chap. Introspection, pp. 49–65.

Vice, Samantha (2006). "Living with the Self: Self-Judgement and Self-Understanding". In: *Judging and Understanding: Essays on Free Will, Narrative, Meaning and the Ethical Limits of Condemnation*. Ed. by Pedro Alexis Tabensky. Burlington, USA: Ashgate Pub Co.

Vice, Samantha (2007). "The Ethics of Self-Concern". In: *Iris Murdoch. A Reassessment*. Ed. by Anne Rowe. New York: Palgrave Macmillan, pp. 72–84.

Walesh, Stuart G. (2016). *Write to Find Out What We Are Thinking and To Learn*. URL: `http://www.helpingyouengineeryourfuture.com/write-to-find-out.htm`.

Watson, Gary (1975). "Free Agency". In: *Journal of Philosophy* 72.8, pp. 205–220.

Wegner, Daniel M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT press.

Weintraub, Ruth (1987). "Unconscious Mental States". In: *The Philosophical Quarterly* 37.149, pp. 423–432.

White, Peter A. (1980). "Limitations on Verbal Reports of Internal Events: A Refutation of Nisbett and Wilson and of Bem". In: *Psychological Review* 87.1, pp. 105–112.

— (1988). "Knowing more about what we can tell: "introspective access" and causal report 10 years later". In: *British Journal of Psychology* 79.1, pp. 13–45.

— (1989). "Evidence for the use of Information about internal events to improve the accuracy of causal reports". In: *British Journal of Psychology* 80.3, pp. 375–382.

Williams, Bernard (1973). *Problems of the Self*. New york: Cambridge University Press.

— (1981). *Moral Luck*. New York: Cambridge University Press.

— (1985). *Ethics and the Limits of Philosophy*. Cambridge, Massachusetts: Harvard University Press.

— (2000). *Morality, an Introduction*. Cambridge: Cambridge University Press.

— (2002). *Truth and Truthfulness*. Princeton, New Jersey: Princeton University Press.

Wilson, Timonthy D. and Simine Vazire, eds. (2012). *Handbook of Self-knowledge*. New York: The Guilford Press.

Wilson, Timothy D. (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Wilson, Timothy D. and Elizabeth W. Dunn (2004). "Self-knowledge: Its Limits, Value, and Potential for Improvement". In: *Annual Review of Psychology* 55, pp. 493–518.

Winnicott, D. W. (1965). "Ego Distortion in Terms of True and False Self'. In: *The Maturational Process and the Facilitating Environment: Studies in the Theory of Emotional Development*. New York: International UP Inc., pp. 140–152.

Winter, Michael Jeffrey (2012). "Does Moral Virtue Require Knowledge? A Response to Julia Driver". In: *Ethical Theory and Moral Practice* 15.4, pp. 533–546.

Wolf, Susan (1986). "Self-Interest and Interest in Selves". In: *Ethics* 96.4, pp. 704–720.

Wollheim, Richard (1984). *The Thread of life*. Cambridge, Massachusetts: Harvard University Press.

— (1991). *On The Emotions*. New Haven: Yale University Press.

— (2003). "On the Freudian Unconscious". In: *Proceedings and Addresses of the American Philosophical Association* 77.2, pp. 23–35.