On the Scaffolding of Selfhood: Self-Representation and the Limits of Misidentification

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:27007740">http://nrs.harvard.edu/urn-3:HUL.InstRepos:27007740</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
# Table of Contents

Title Page 1

Table of Contents 2

Acknowledgments 3

Preface 4-5

Glossary of Abbreviations 6

Abstract 8-9

Introduction
  
  *Ia. Background: State of the Field and Literature Review* 10-12
  
  *Ia.1: Nosology* 12-13
  
  *Ia.2: Phenomenology and descriptions of disorder* 13
  
  *Ia.3 Metaphysics and Ontology* 13-14
  
  *Ia.4: Interdisciplinary Typology* 14-16

*Ib. Purpose and Significance of Present Inquiry* 17-18

Methods 19

Results & Discussion
  
  I. Pathologies of Thought and First Person Authority 20-37
  
  II. Mirror Synesthesia and the Limits of Misidentification 38-50
  
  III. Constructing the Essential Indexical ‘I’ 51-70

Future Directions 71-72

Summary and Conclusions 73-74

References 75-83

Appendices 84-97
Acknowledgements

I am immensely grateful to the numerous individuals who have generously fostered and guided my interest and research in philosophy of medicine, philosophy of mind and medical ethics over the past four years. Dr. Harold Bursztajn has served as a mentor since my first semester in Harvard Medical School. Beyond the myriad lessons he has taught me about philosophy in and of medicine, and about the relevance of philosophical inquiry to the foundations and practice of neuropsychiatry and bioethics, he has served as a formative model of wisdom who has deeply impacted the kind of clinician and thinker that I aspire to be. Throughout all stages of the present thesis he has provided exceptional guidance and has always been generous with his time and thoughts despite his busy schedule.

I also owe deep thanks to Dr. Martin Samuels, Dr. Scott Podolsky, Dr. Kirk Daffner (Harvard Medical School) and Gilad Bendheim (Harvard Law School) for insightful comments and conversations during nascent stages of this project; to the Petrie-Flom Center for Health Law Policy, Biotechnology and Bioethics, which supported my work on foundational issues in bioethics through a fellowship during my second year of medical school; to the Harvard University Philosophy Department, who graciously enabled my access to their extensive philosophy library collection during weekends; Drs. Lisa Lehmann (Harvard Medical School), Scott Halpern (University of Pennsylvania), Stephen Yalowitz (University of Maryland) and Arthur Caplan (NYU) whose early guidance and examples have served to animate and shape my interest in and approach to philosophy of medicine and bioethics; and to my parents and grandparents, who have instilled me with ideas and ideals that have informed and inspired me in immeasurable ways.
Preface

My involvement in philosophy of medicine and bioethics has spanned the entirety of my medical school career; during this time, I have led and contributed to research projects in this field, resulting in the following publications:

1. Young MJ. “Pathologies of Thought and First-Person Authority.” Philosophy, Psychiatry & Psychology. Forthcoming.

This thesis focuses on an investigation of selfhood through the specific clinical lenses of mirror touch synesthesia and thought-insertion delusions, analyses that are in part forthcoming in Philosophy, Psychiatry & Psychology (#1-2 above); to be published in that same volume are
commentaries by Dr. Lauren Ashwell (Bates College) and Dr. Clara Humpston (Cardiff University) on my work on thought-insertion delusions, to which my analysis in #2 responds *inter alia*, and which are included in the appendix of this thesis for readers’ reference. Work from my other publications is drawn upon in the introduction, discussion, and concluding sections. I was involved in and led all aspects of the present investigation including structural design, analysis, manuscript preparation, submission for peer review, and revision. The contributions of many others are outlined in the above Acknowledgements section.
Glossary of Abbreviations

AD: Alzheimer’s disease
CPR: Kant’s *Critique of Pure Reason*
DSM: Diagnostic and Statistical Manual for Mental Disorders
FTD: Frontotemporal dementia
TI: Thought-insertion
IEM: Immunity to error through misidentification
IP: Immunity principle
PI: Wittgenstein’s *Philosophical Investigations*
And out of what one sees and hears and out
of what one feels, who could have thought to make
so many selves, so many sensuous worlds...

- Wallace Stevens, Esthétique du Mal (XV)
Abstract

Insofar as psychiatrists and neurologists tend to the mental wellbeing of others, their work is interwoven with philosophical concerns and theoretical assumptions about the nature of the mind, its myriad functions and the conditions governing its multiform pathologies. That the mind figures so prominently in their ordinary language attests to the wealth of insights that stands to be gained through a dialogue with philosophy. In one of the earliest efforts to taxonomize psychiatric medicine, General Psychopathology, Jaspers incisively remarks that “the exclusion of philosophy would be disastrous for psychiatry...if any psychopathologist thinks he can disregard philosophy and leave it aside as useless he will eventually be defeated by it in an unperceived way” (770). At the very least, philosophy can offer psychiatry, neurology, and indeed all of medicine a more refined vocabulary to describe the human phenomena it aims to capture. Adding to Jaspers’ remarks, it is important to appreciate the significance of the reverse as well: a philosophical enterprise that confines itself to its own increasingly scholastic linguistic games and overlooks the findings of neurology, psychiatry and the neurosciences will eventually be overwhelmed by them in some form or other.

These considerations animate the present work, which examines the crucial yet underexplored convergence of philosophy and neuropsychiatry with respect to self-representation and first-person authority, particularly through the lens of thought-insertion delusions and mirror synesthesia, and illuminates a novel account of selfhood emerging at this intersection.

After an introduction describing the state of the field of philosophy of medicine and philosophy of neuropsychiatry, Chapter 1 – Pathologies of Thought and First-Person Authority, critically examines how the psychopathology of thought-insertion delusions can inform our philosophy of mind, and how philosophical insights may in turn help us describe the nature of thought-insertion. Some philosophers have argued that the phenomenon of thought-insertion challenges conventional assumptions about first-person authority, and in particular, the immunity principle (IP), according to which first-person, present-tense reports about psychological states are immune to errors of identification. This chapter elucidates different ways that the immunity principle can be set up and examine what thought-insertion delusions tell us about the sustainability of each. By way of this analysis, I show why it is crucial that any formulation of the immunity principle be limited to present-tense self-referential judgments (in particular, judgments ascribing some predicate to the essential indexical ‘I’) that are based on a certain kind of personal perspective. Based on this understanding, it is argued that that the phenomenon of thought-insertion (TI) poses no real problem for the immunity principle. The craving for more philosophical generality with respect to this principle, I contend, has led various theorists to oversimplify the principle and to think that they must jettison it when confronted with deviant psychological phenomena such as thought-insertion delusions. The structure of this chapter is as follows: §II explains why self-knowledge is distinctive and examines what it means for judgments based on self-knowledge to be immune to errors through misidentification with respect to the first-person. §III introduces the phenomenon of thought-insertion and presents Campbell’s argument that this phenomenon challenges the assumption that first-person, present-tense reports about psychological states are immune to errors of identification. §IV explains why the difficulties raised by Campbell are blocked by the analysis developed in §II, and argues that the
explanatory moves some make to demystify mysterious aspects of TI unnecessarily distort its apparent phenomenology. §V examines why TI remains puzzling, even if it does not challenge any obvious philosophical principle.

Chapter 2 – Mirror Synesthesia and the Limits of Misidentification, explains and critically evaluates arguments raised in response to my accounts of IP and TI. §I offers a rejoinder to the critique that IP does not turn on features unique to the first-person, and argues that the epistemic gap between suggestion and indefeasible demonstration with respect to evidence opens up the possibility for an agent to justifiably take as x as evidence for p in situation z even if it is the case that ~p in z.1 §II responds to the strategy of bisecting TI into an inserted-thought and a false belief about the inserted-thought, argues that a more plausible conception of thought-insertion delusions may be non-doxastic, and explains the compatibility of either account with IP. §III examines possibilities for non-delusional first-person-like awareness of mental states that are not one’s own. The condition of mirror synesthesia is introduced, and I argue that instances of mirror synesthesia blur the boundaries between identification and misidentification, and challenge philosophical orthodoxies about the self in a manner intriguingly similar to TI.

Chapter 3 - Constructing the Essential Indexical, situates foregoing insights about self-other representation into a broader account of selfhood. §I explains that the capacity for self-other representation dealt with in Chapters 1 and 3 is not wholly constitutive of selfhood. §II delineates and defines other capacities and functions necessary for an intact scaffolding of selfhood. §III examines how substantial alterations of any of these elements of selfhood could result in a fracturing, disintegration or transformation of the emergent self, resulting in an array of conditions from the ordinary (e.g., conventional akasias) to the pathological. On the basis of this analysis, it encourages future appraisal of how particular lesions may differentially affect elaborations of selfhood by impacting neuronal substrates corresponding to different dimensions of this complex scaffolding, while emphasizing caution with respect to tempting mereological fallacies in this domain. It concludes by illuminating fruitful areas for future research, and by urging sustained collaboration between philosophers and physicians to optimize and advance approaches to the diagnosis, evaluation and treatment of pathologies of the mind involving disruptions of selfhood.

1 The tilde symbol (~) used throughout denotes strict negation in concordance with the conventions of first-order predicate calculus; ‘~T’ should therefore be read as ‘it is not the case that T’, where ‘T’ is a placeholder for any truth-functional proposition. The horseshoe symbol (⊃) is used throughout to denote material implication; ‘P ⊃ Q’ should therefore be read as ‘if P then Q’, which is logically equivalent to ‘~(P&~Q)’, where ‘P’ and ‘Q’ are placeholders for any truth-functional propositions.

I. Introductory Remarks

Ia. Background: State of the Field and Literature Review


In his influential 1976 essay on the relationship of philosophy with medicine, the late Dr. Edmund Pellegrino elegantly described that “[m]edicine and philosophy oscillate about each other like the strands of a complex double helix of the intellect. They are intermittently drawn together by their immersion in man's existence and driven apart by their often opposing preoccupations with that existence. Special tensions arise from their conflicting claims to universality—medicine divinizing the body and the particular, philosophy the intellect and the abstract. The state of the relationship is, however, always of profound cultural significance. When medicine and philosophy converge, they can greatly advance man's search for a unified image of himself and the world; when they diverge, that image becomes fragmented, puzzling, and even absurd.” The vision Pellegrino articulates underscores the wealth of insights that can be afforded by the analytic alliance of philosophy and medicine.

Beyond symbiotic content, philosophy and medicine display a distinctive methodological kinship. This kinship can be found gestured toward in the later work of Ludwig Wittgenstein, who in his 1953 *Philosophical Investigations* (PI) incisively observed that “[t]he philosopher’s
treatment of a question is like the treatment of an illness” (§255), and correspondingly, “there is not a philosophical method, though there are indeed methods, like different therapies (§133). Wittgenstein accordingly makes regular habit of referring to foundational philosophical errors as ‘diseases’, as for instance in the Brown Book where he describes “[a] kind of general disease of thinking which always looks for (and finds) what would be called a mental state from which all our acts spring, as from a reservoir” (143), and elsewhere in PI where he notes “[a] main cause of philosophical disease - one-sided diet: one nourishes one's thinking with only one kind of example” (§593).

Arguments for the analytic alliance of philosophy and medicine trace back to Galen (c. 130 – 200 CE) whose treatise Quod optimus medicus sit quoque philosophus (“That the Best of Physicians is also a Philosopher”) describes the importance of natural philosophy, logic and ethical theory to the ideal practice of medicine. Galen’s influential approach motivated the incorporation of philosophical training into late Alexandrian and medieval Arabic medical curricula², and informed many later physician-philosophers such as Maimonides, Descartes, John Locke and Francisco Sánchez, who describes at the outset of his 1581 treatise on knowledge and the scientific method, Quod Nihil Scitur, that “[t]he goal of my proposed journey is the art of medicine, which I profess, and the first principles of which lie entirely within the realm of philosophical contemplation.” And during the Scientific Revolution, as Steven Shapin (2000) intriguingly notes, an “important gauge of the quality of reformed natural philosophical knowledge was its ability to produce a more effective medical practice,” (131) with Descartes making it his explicit aim in his 1637 Discourse on Method to “devote the rest of my life to nothing other than trying to acquire some knowledge of nature from which we may derive rules of medicine which are more reliable than those we have had up till now” (136).

Few areas of medicine encapsulate as concentrated and colorful populations of philosophical issues as psychiatry and neurology, specialties that share their primary organ of

interest with a major branch of philosophy - philosophy of mind. Research at this interface is characterized by at least four distinctive streams, each of which carries unique theoretical and practical utility. Each of these four streams is reviewed briefly below.

**Nosology**

At the level of disease categorization and description, there is a stream of research that examines the philosophical underpinnings of and justification for psychiatric and neurologic nosology. Research in this stream can be either descriptive or normative in its aim; that is to say, theoreticians might embark on the project of illuminating the nature of a given nosology with the aim of characterizing its present state, function and implications, or with the aim of devising a more conceptually sound approach to constructing an operative taxonomy of disease. Some have argued that the relevance of such nosologic issues is generally greater with respect to brain disorders than in general medicine, owing in large part to the relative paucity of comprehensive and transferable mechanistic knowledge about brain disorders; as a result, diagnostic terms for brain disorders are often formulated around clusters of symptoms rather than around knowledge of causal architecture (Cf. Graham & Stephens 1994; Maung 2016; Sisti, Caplan & Young 2013). Philosophical efforts in this area of nosology are especially relevant to help provide conceptually clarity in times such as ours where burgeoning knowledge of the brain and its development has prompted a flurry of uncertainty among practitioners and researchers surrounding whether and how diagnostic terms ought to be revised to reflect emerging connectomic and genomic discoveries (Insel et al. 2010; Klaas et al. 2016).

**Phenomenology and Descriptions of Disorder**

Second, on a phenomenological plane, there is a stream of research and practice that wields analytic tools of philosophy of mind and existentialism to advance enriched understandings of and therapeutic approaches toward patient presentations and apparent experiences. Examples of work falling within this stream include Leston Haven’s existential psychotherapy, Karl Jaspers phenomenological methodology and Eugene Minkowski’s approach to analysis, among many others, heavily influenced by the existentialist philosophy of Sartre and transcendental phenomenology of Husserl (Beveridge 2011; Cf., Bursztajn 1972). Jaspers
incisively characterizes the approach as giving “concrete description[s] of the psychic states which patients actually experience and presents them for observation. It reviews the inter-relations of these, delineates them as sharply as possible, differentiates them and creates a suitable terminology... Detached psychiatric observation with its own formulation of what the patient is suffering is not any substitute for this” (GP, 55, my emphasis). Additionally, many neuroscience researchers, most notably those involved in brain-cognition mapping, have recently displayed renewed interest in phenomenology, with some advocating for “neurophenomenology...which uses phenomenological information about the variability of subjective experience in order to illuminate the variability of brain dynamics, offers a promising method for addressing this challenge” (Fazelpour & Thompson 2015: 223).

In elucidating the structure of conscious experiences as given by the first-person perspective, such inquiry provides distinctive insight into forms of life – what Wittgenstein dubbed Lebensformen (and, though not identical, in Husserlian terminology Lebenswelt), and which in his view held the key to solving many philosophical problems which arose not by virtue of yet undiscovered metaphysical truths but rather by virtue of a lack of sufficient attention to what is already given in lived experience and by language in use.

Hence, submits Wittgenstein, “the first step [in the generation of a philosophical problem] is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometimes perhaps we shall know more about them – we think. But that is just what commits use to a particular concept of what it means to learn to know a process better. (The decisive moment in the conjuring trick has been made, and it was the very one we thought quite innocent)” (PI §308). Echoing the phenomenological method, he contends that “[t]here must not be anything hypothetical in our considerations... descriptions alone must take its place...These are, of course, not empirical problems, they are solved, rather, by looking into the workings of our language, and that in such a way as to make us recognize those workings: in despite of an urge to misunderstand them. The problems are solved, not by giving new information, but by arranging what we have always known... What we do is to bring words back from their metaphysical to their everyday use... It earmarks the form of account we give, the way we look at things. (Is this a 'Weltanschauung'?)” (PI §108; 116; 122). Given the rich phenomenological content of clinical encounters, insights gleaned by way of this method in medical contexts are as
relevant to theoreticians (who may look toward the clinic as a conceptual laboratory, as it were) as they are for practitioners, who through this method may identify core features of patient experiences to be monitored and treated.

**Metaphysics and Ontology**

Third, there is a stream of research that wields discoveries in neuroscience and empirical neuropsychiatric data to address questions in metaphysics, spanning topics including of free-will and determinism (Cf. Wegner 2003; Tse 2013); personal identity (Cf. Klein 2013; Strohminger & Nichols 2015; Fuchs 2016); the mind-body problem (Hacker 2015), akrasia and moral responsibility (Cf. Baum 2016, Yaffe 2013); and consciousness (Cf., Block 2015, Navajas & Kaunitz 2016). Despite significant theoretical overlap in conceptual territory between philosophy of mind and neuroscience, many issues and topics in at this interface remain critically underexplored, and it is the exception rather than the norm for neuroscience researchers or practitioners to consult or collaborate with philosophers of mind or vice versa in the course of their investigations.

**Interdisciplinary Typology**

At a higher level of generality, a fourth stream of research at this interface examines the topology of conceptual overlap and general dynamics of interplay between the fields of philosophy, psychiatry and neurology. A paradigmatic instance of this sort of inquiry can be found in Dr. Edward Hundert’s 1989 book *Philosophy, Psychiatry and Neuroscience: Three Approaches to the Mind*, wherein he critiques “the traditional separation of philosophy, psychiatry and neuroscience into distinct academic disciplines [which results in] discrete and unrelated approaches to the mind,” and advocates a unified theory wherein contemporary “psychiatry and neurobiology can be understood as a natural continuation of Hegel’s nineteenth century critique of Kantian philosophy.”

Beginning his book with an incisive quote from Brain’s Clinical Neurology “[t]he brain and the mind constitute a unity, and we may leave to the philosophers, who have separated them in thought, the task of putting them back together again” (479), Hundert argues that “the insights gained from multi-disciplinary synthesis will leave the burden of proof...with those purists who
claim that these contingent matters are ‘not the philosopher’s business’” (303). He concludes that “[t]he only way philosophy will further uncover the world’s metaphysical laws of working through the study of human experience is by abandoning all forms of dogma and opening itself wider to the truth found in the actualities of our immediate, lived experience...This is, of course, the traditional way philosophy progresses...” (304).

Though not commenting on the relationships between philosophy and these areas of medicine in particular, Cambridge philosopher Jane Heal underscores the importance of this kind of interdisciplinary inquiry in identifying a perpetual risk for investigative enterprises writ-large to lose touch with their original “nourishing roots in the concerns of human life in general” (Heal 40). What gets an investigative pathway or entire discipline going, Heal notes, “may be problems, puzzlements, uneasinesses, and half-grasped insights... But as such discussion develops it may lose touch with its original motivations in life outside itself and fail to develop replacement roots. And when that happens, assessment of the potential merits or demerits of any new moves that may be offered as the enterprise advances will come to be assessed under a hodgepodge of criteria, which answer neither to the problems, puzzlements, and so on, that were at the origins of the thing nor to problems, puzzlements, and so on, in contemporary life outside the practice, but are instead internally generated and hence at the mercy of fashion, personalities, or institutional [idiosyncrasies]... institutions which encourage wide and outward-looking intellectual sympathies, and which do not reward narrow point-scoring specialism, are helpful in resisting the tendency to scholasticism... [hence, many] philosophers resisted narrowing professionalism within the subject. They worked in a way which presupposed that its various strands, for example logic and ethics, could and should influence eachother....Plainly Wittgenstein comes to mind as a prime example... [such thinkers were] generalists, not in the sense that they dabbled in many different things, but in the sense that they tried to say things that were of more than parochial or specialist interest.”

Most recently, Dr. Kenneth Fulford and others have continued in this vein, explaining the indispensability of philosophical tools in coherently optimizing approaches to patient care, and developing a clinical system dubbed “values-based practice” (VBP) which advocates for explicit recognition of different values and philosophical principles at play in clinical encounters,
and arguing that recognition of philosophical superstructure is critical in crafting more effective models of service delivery, research and education.

The four streams of inquiry described above are intended to provide the reader with a logically organized sense of the breadth and background of the state of the field; it is not meant to imply that these streams of inquiry are disconnected from one another, or, taken in sum, exhaustive of the conceptual territory. On the contrary, each of these streams of inquiry crisscrosses and ineluctably informs one another.
Ib. Purpose and Significance of the Present Inquiry

In keeping with the foregoing schema, the inquiry pursued in the present thesis is situated within streams two and three as articulated above. Its general aim is to examine the interface of philosophy and neuropsychiatry with respect to selfhood and first-personal experience, and explores this area specifically through the clinical lens of mirror-touch synesthesia and delusions of thought-insertion. In short, the question it addresses can be put thus: what does neuropsychiatry (in particular, thought-insertion delusions and mirror synesthesia) reveal about the scaffolding of selfhood, and how can philosophy help elucidate the phenomenology of disorders in which this scaffolding is apparently disturbed?

Examining these issues, I argue, sheds much light on philosophical, clinical and neuroscientific conceptions of selfhood and its attendant ontological and epistemological features. In addition to illuminating novel epistemological points about privileged access and the limits of misidentification, it also, I argue, furnishes clinicians with better understandings of the experiences of patients who suffer disturbances of selfhood.

Of note, the approach one takes to these issues will bear on how one understands selfhood, action, agency, and their proximal source - the mind. In turn, these insights portend significant practical consequences; in addition to furnishing more accurate criteria for describing and diagnosing pathologies of selfhood, such understandings can influence (i) how and when responsibility, blame and culpability could be aptly assigned to agents with putative disturbances of selfhood and (ii) how, if at all, pathologies of selfhood can be optimally prevented, treated or palliated. If the operative account of selfhood is flawed, research and clinical strategies developed thereon will reflect this. In the hands of policy makers, neuropsychiatrists, psychologists, economists, philosophers, judges and juries, different approaches to selfhood can give rise to radically different conclusions and styles of practical guidance. As we will see, in

---

the hands of surrogates, how to understand and relate to one who experiences a disturbance of selfhood (as, for example, in schizophrenia or advanced Alzheimer’s disease) can be swayed by what approach to selfhood is taken; indeed, it can shape how the patient views their own condition. Significant practical and theoretical issues therefore hinge on the present project.

Methods

The methods employed by this project derive from the traditions and techniques of modern analytic philosophy of mind, epistemology and philosophy of language.
Chapter 1:
Pathologies of Thought and First-Person Authority

I. Introduction

Insofar as psychiatrists and neurologists tend to the mental wellbeing of others, their work is interwoven with philosophical concerns and theoretical assumptions about the nature of the mind, its myriad functions and the conditions governing its multiform pathologies. That the mind figures so prominently in their ordinary language attests to the wealth of insights that stands to be gained through a dialogue with philosophy. In one of the earliest efforts to taxonomize psychiatric medicine, General Psychopathology, Jaspers incisively remarks that “the exclusion of philosophy would be disastrous for psychiatry…if any psychopathologist thinks he can disregard philosophy and leave it aside as useless he will eventually be defeated by it in an unperceived way” (770). At the very least, philosophy can offer psychiatry (and all of medicine for that matter) a more refined vocabulary to describe the phenomena it aims to capture. Adding to Jaspers’ remarks, it is important to appreciate the significance of the reverse as well: a philosophical enterprise that confines itself to its own increasingly scholastic linguistic games and overlooks the findings of neurology, psychiatry and the neurosciences will eventually be overwhelmed by them in some form or other.

This chapter critically examines how the psychopathology of thought-insertion delusions can inform our philosophy of mind, and how philosophical insights may in turn help us describe the nature of thought-insertion. It has been argued that the phenomenon of thought-insertion challenges conventional assumptions about first-person authority, and in particular, the immunity principle (IP), according to which first-person, present-tense reports about psychological states are immune to errors of identification (Campbell 1999). Here I elucidate different ways that the immunity principle can be set up and examine what thought-insertion delusions tell us about the sustainability of each. By way of this analysis, I show why it is crucial that any formulation of the immunity principle be limited to present-tense self-referential judgments (in particular, judgments ascribing some predicate to the essential indexical ‘I’) that are based on a certain kind
of personal perspective. Based on this understanding, I argue that that the phenomenon of thought-insertion poses no real problem for the immunity principle. The craving for more philosophical generality with respect to this principle, I argue, has led various theorists to oversimplify the principle and to think that they must jettison it when confronted with deviant psychological phenomena such as thought-insertion delusions.

§II explains why self-knowledge is distinctive and examines what it means for judgments based on self-knowledge to be immune to error through misidentification with respect to the first-person. §III introduces the phenomenon of thought-insertion and presents Campbell’s argument that this phenomenon challenges the assumption that first-person, present-tense reports about psychological states are immune to errors of identification. §IV explains why the difficulties raised by Campbell are blocked by the analysis developed in §II, and argues that the explanatory moves some make to demystify mysterious aspects of TI unnecessarily distort its apparent phenomenology. §V examines why TI remains puzzling, even if it does not challenge any obvious philosophical principle.

II. Self-knowledge and Immunity to Error

Orthodoxy about self-knowledge tells us that certain thoughts that we have about ourselves are made available to us in a distinctive, authoritative way (e.g. Wittgenstein 1958,66-67, Shoemaker 1968, Davidson 1984, Campbell 1999). That is to say, there is a unique way of knowing is not ordinarily available to me when I try to acquire thoughts about other persons or objects. The way I come to know ‘I am in pain’, for instance, is different from the way I ordinarily come to know ‘my mother is in pain’. This distinctive way of knowing has often been taken to characterize judgements about one’s thoughts or psychological properties, but also judgements made in the normal way about one’s bodily positions (e.g., ‘my arms are crossed’) (Evans 1982,219-221). Here I will be dealing mainly with the former, for it is judgements of this sort that are at stake in the cases of thought-insertion that we will discuss. It is not the case that every type of thought that one has about oneself is obtained via this distinctive way of knowing. For instance, one may think ‘I am anemic’ because the blood test revealed a low serum platelet
count. Rather there is a more limited set of token thoughts that are obtained through a unique first-person perspective. I will hereafter refer to token thoughts obtained by way of this privileged first-person perspective as thoughts*. My aim in the present section is not to explicate the mechanism(s) by which this way of knowing operates; rather, I take it that this way of knowing exists in some form as a given and describe how the thoughts it produces are distinctive. With this understanding, we will be in a better position to evaluate the phenomenon of thought-insertion and Campbell’s argument.

Thoughts* share at least two distinctive features. First, they are known to one immediately, without the need for any external evidence, investigation, or complex inference (Davidson 1984). This is not to say that such thoughts are infallible, but rather that the way in which they come about is such that the sort of evidence or inference that would typically be needed if the thoughts did not come about in this first-person way is not needed when they do come about in this way. When I walk through the snow and think, in the usual sort of way, ‘I feel cold’ or ‘my arms are crossed’, this is known to me immediately; I do not need to take my own temperature or to confirm that it really is me who I identify as feeling cold or whose arms are crossed. In such cases, the very same information that conveys to me that something instantiates a given property immediately conveys to me who it is that instantiates that property. This is not to say that no evidence at all is needed for such proprioceptively/introspectively grounded judgments, but rather that if it is the case that any such evidence is needed then that evidence is immediately available to me. For this reason, I have chosen to refer to evidence that is not immediately available to me as ‘external evidence’. By ‘immediately available’ I mean discernible to me solely on the basis of introspection, proprioception or some other form of direct inner awareness, in contradistinction to thoughts that I come to form on the basis of observation or interaction with the outside world.

Second, while thoughts* are not infallible, they are distinctive in the sense that they seem to be immune to errors of misidentification with respect to the first-person. What are errors of misidentification, and what does it mean for thoughts* to be immune to them in this way?

To grasp what is meant by ‘errors of misidentification’, suppose you are gazing out the window and you think that you see your friend Bob wearing a wool hat. You subsequently form the following judgment:
(1) Bob is wearing a wool hat.

There are two key ways in which you can go wrong with respect to this judgement. First, you might be mistaken about whether it is indeed a wool hat that Bob is wearing (it might be cotton, or not a hat at all). This kind of error is an error of mispredication. Errors of this kind occur when, despite having correctly identified the subject of a judgement, one ascribes to the subject a property it doesn’t have. Alternatively, you may rightly detect the property ‘wearing a wool hat’ but be wrong it is about whether it is Bob who is wearing it. This kind of error is an error of misidentification. Errors of misidentification occur when, despite having correctly identified a property, one misidentifies the subject instantiating that property.

So, for any agent A, a thought of the form ‘X is F’ is *immune to error through misidentification* with respect to the referring place here occupied by X just in case the following is impossible:

(EM) A is right on the basis of the perspective grounding his thought that something is F but wrong about whether X is F.

Token thoughts that we acquire through our distinct first-person way of knowing (i.e., thoughts*) appear to be immune to this kind of error with respect to the first-person. The idea here is that if the indexical referring expression ‘I’ is used sincerely and understandingly in a thought* of the form ‘I am F’ then that person cannot be right on the basis of the perspective that generates this thought that someone or something is F but wrong about whether ‘I am F’. So, for instance, the thought*

(2) I am cold

is immune to error through misidentification, since I cannot be right on the basis of the perspective grounding this thought* that *someone* is cold but wrong about whether I am the one who is cold.
By contrast, suppose that after reading the doctor’s report I form the thought (that is, as seen earlier, not a thought*)

(3) I am anemic

This thought would not be immune to error through misidentification, for I could be correct on the basis of the perspective grounding this thought that someone is anemic but wrong about whether I am the one who instantiates the property of being anemic. For instance, the doctor may have mistakenly sent the report of a different person with a similar name to me. This thought is susceptible to errors of misidentification in the same way that (1) is. It thus seems that in order for a token thought to be immune to error through misidentification with respect to the first-person it must be based on the distinctive kind of first-person perspective that is characteristic of thoughts*.

The foregoing notion will hereafter be referred to as the Immunity Principle, which is intended as a definition of ‘immunity to error through misidentification’:

(IP) any agent who thinks a thought* of the form ‘I am F’ cannot be right on the basis of the perspective grounding this thought that something is F but wrong about whether ‘I am F’.

The 'on the basis of the perspective grounding this thought' clause in (IP) is essential, for if we just had

(IP’) Any agent who thinks a thought* of the form ‘I am F’ cannot be right that something is F but wrong about whether ‘I am F’

then the truth of any token thought* ‘I am F’ made by any agent a would be guaranteed so long as a knew of some x that Fx. For instance, (IP’) would guarantee the truth of delusional thoughts* such as ‘I am dead’ so long as the person who has that thought* knows that there is
someone who is dead. This clause thus guarantees that the claims immunized by IP are claims involving properties that are correctly detected in the right kind of way. The same perspective that gives the agent the information that there is some property F must give him the information that he is the instantiator of that property in order for the corresponding thought ‘I am F’ to be immune to errors of misidentification.

Moreover, we avoid

(IP’’) Any agent who thinks a thought* of the form ‘I am F’ cannot be right on the basis of this thought* that something is F but wrong about whether ‘I am F’.

for this would leave us with the unpleasant problem of explaining how a thought can be based on itself.

To be sure, judgments that are immune to errors of misidentification need not be first person judgements. As Shoemaker (1968) suggests, judgments featuring other indexical referring expressions may also be immune to errors of misidentification (i.e. ‘this is red’, ‘that is loud’). Judgments of this sort that are immune to errors of misidentification are similar to I-thoughts* in that the information that makes it sensible for one to think that there is some property being instantiated concomitantly individuates the thing that instantiates the property. By contrast, judgments that are susceptible to errors of misidentification rely on some distinct identity judgement that can go awry. (3), for instance, relies on the defeasible identity judgement (‘I’ = ‘the person the doctor’s report is about’). Likewise (1) relies on the identify judgement (‘the person over there’ = ‘Bob’). Intuitively, the fact that thoughts* are obtained by way of the distinctive first-person way of knowing seems to obviate the need for distinct identification of the instantiator of the property discerned. For the remainder of this chapter I deal primarily with thoughts* that are immune to errors of misidentification with respect to the first-person, for these are what are brought into question by the phenomenon of thought-insertion.
III. Thought-insertion: A Challenge to the Immunity Principle?

Experiences of thought-insertion (TI) are first-rank symptoms of schizophrenia wherein patients report having thoughts that are not their own. Inserted thoughts are reported as being completely alien - not just as being *caused* by another agent but as *belonging* to another agent. In this respect, delusions of thought-insertion can be differentiated from delusions of control in which one thinks something like ‘my thoughts are being controlled by aliens’.

Consider the following case reports:

Thoughts come into my mind like ‘kill God’ [(henceforth ‘KG’)]. It is just like my mind working, but it is not. They come from this chap, Chris. They're his thoughts (Frith 1992,65).

[T]he feeling suggests that somebody has attached himself to my mind and feeling, just as in a game of cards someone looking over one’s shoulders may interfere in the game (Jaspers 1963,123).

…[T]hey come unasked... They come at any moment like a gift and I do not dare impart them as if they were my own (Jaspers 1963,123).

[S]he said that sometimes it seemed to be her own thought ‘but I don’t get the feeling that it is...the feeling is not the same...the feeling is that it is somebody else’s.’ (Hoerl 2001,190).

In each of these cases, one encounters a thought from a first-person perspective but immediately experiences the thought as belonging to another. As Jaspers explains, in such cases “[p]atients think something and yet think that someone else has thought it...the thought arises and with it a direct awareness that it is not the patient but some external agent who thinks it.” (1963,122).
Campbell (1999) argues that cases of this kind give us reason to reject the conventional notion that first-person, present-tense reports of psychological states are authoritative. He writes that

“What is so striking about the phenomenon of thought-insertion as described by schizophrenic patients is that it seems to involve an error of identification…A patient who supposes that someone else has inserted thoughts into his mind is right about which thoughts they are, but wrong about whose thoughts they are. So thought-insertion seems to be a counterexample to the thesis that present-tense introspectively based reports of psychological states cannot involve errors of identification” (609).

Consider, for instance, the case cited by Frith above. Through his first-person privileged way of knowing, patient A encounters the thought ‘KG’. He disavows this thought as his own, and claims that it comes from ‘this chap, Chris.” In this case, the subject misidentifies whom his mental state belongs to, despite the fact that he appears to encounter this thought in the distinctive, first-person way.

Campbell draws from these cases that “experience[s] of thought-insertion [exhibit] error[s] of identification, though it is sometimes taken to be a logical point that judgments about one’s own current thoughts are immune to errors of identification” (619-620).

From Campbell’s point of view, this leaves us with the option of conceding that the logical principle exists but denying that the thoughts of the schizophrenic have stable meaning, or ascribing stable meaning to the thoughts of the schizophrenic but sacrificing the supposed logical principle. Given that these thoughts do have some kind of meaning (as we do seem to understand them), he is inclined to think we should deny the supposed logical principle. But might there be a way through the horns of this dilemma, to preserve the principle without denying that such thoughts have meaning?
IV. Thought-insertion: Rethinking the Challenge

While cases of thought-insertion might at first glance appear to challenge our assumptions about the immunity of thoughts* to errors of identification, when we look more closely it is unclear how the putative counterexample is supposed to work. What exactly is the thought* that Campbell thinks is problematic, and what is the “logical point” at stake?

One possibility is the following thought formed by patient A:

(4) Chris is the one who thinks ‘KG’

Even if Campbell were granted the assumption that (4) is a thought* in the first place, in order for (4) to be rightly regarded problematic, one would have to gloss over the fact that thoughts* are only immune to errors of misidentification with respect to the first-person. So, if (4) is what Campbell has in mind as his counterexample, then the “logical point” he refers to must be something like

(IP’’’) any agent A who thinks a thought* of the form ‘X is F’ cannot be right on the basis of the perspective grounding this thought that something is F but wrong about whether ‘X is F’.

This rendering follows Campbell’s understanding that "present-tensed introspectively based reports of psychological state cannot involve errors of identification" (Campbell 1999:610). If it is the case that patient A thinks that ‘Chris is the one who thinks ‘KG’’ in the privileged, first-person way, then this would indeed seem to challenge (IP’’’). In this case, patient A thinks a thought* of the appropriate form (i.e. ‘Chris is the one who thinks ‘KG’’); is right that there is someone who thinks ‘KG’ (indeed, the thought exists); but is wrong that ‘Chris is the one who thinks ‘KG’’. Indeed, the thought* is his – it issues from his delusional mind, not from Chris. Ex hypothesi, cases of this kind show us that it is possible for thoughts obtained by way of the distinctive first-person way of knowing to exhibit errors of misidentification. Thus, the argument goes, such cases are counterexamples to the immunity principle.
However, it appears that (IP’’) inadequately expresses the “logical point” at stake, for, as we have seen, thoughts* are only immune to errors of misidentification with respect to the first-person. The idea here is that if the indexical referring expression ‘I’ is used understandingly in a thought* of the form ‘I am F’ then that person cannot be right on the basis of this thought that someone or something is F but wrong about whether ‘I am F’. However, there is no guarantee thoughts* of the form ‘X is F’ are immune to errors of misidentification when ‘X’≠‘I’.

On a careful rendering of the immunity principle, then, a delusional thought such as ‘Chris is the one who thinks ‘KG’’ is not an effective counterexample because the one thinking the thought is not Chris. Thoughts about third parties are not categorically privileged in this way, even if such thoughts are thoughts*.

This illustrates the importance of analytical clarity when referring to the point that “judgments about one’s own current thoughts are immune to errors of identification.” The claim is superficially plausible, but is ambiguous between (IP), (IP’), (IP’’) and (IP’’’). And as we have seen, only (IP) seems to express a solid point.

But perhaps what Campbell has in mind as a counterexample is not (4), but rather patient A’s thought that

(5) I am not a person who thinks ‘KG’

In this case, the patient does appear to use ‘I’ understandingly. Does this challenge our assumption that thoughts* are immune to errors of misidentification with respect to the first-person? Initially, it might seem so. Indeed, one might reason, patient A is the one who thinks ‘KG’, and he is mistaken in thinking otherwise.

One possible response is to deny that such delusional thoughts are of the relevant kind, i.e., to argue that they aren’t obtained through the distinctive, first-person way of knowing and thus are not thoughts* at all. While this approach might evoke some intuitive sympathy, at least two problems render it unattractive. First, it seems like an ad-hoc response to the problem of thought-insertion. Second, and more importantly, it ignores the fact that patients do seem to arrive at their
delusional thoughts through a distinctive, first-person way of knowing. Indeed, these thoughts are immediately available to nobody else other than the patients, and patients seem to require no external evidence, investigation or complex inference in order arrive at them.

However, even if Campbell is right that token thoughts such as (5) are of the relevant kind, it would be misguided to regard them as challenges to (IP) or as involving errors of identification. Here I present two ways of understanding what is going on with such thoughts* such that we can preserve (IP) without denying that such thoughts* have stable meaning.

One strategy is to distinguish varieties of ‘think’ and to interpret the claim of the sufferer that ‘I am not the person who thinks KG’ as involving no error at all. The relationship that one might have to a thought (or the ‘thinking-relationship’) can assume various forms. For instance, a statement of the form ‘I think X’ can mean any of the following: ‘I endorse X’; ‘The thought that X makes the most sense to me’; ‘I am causally responsible for the thought that X’; ‘I am aware of a thought that X’. It is in principle possible that one stand in relation to one’s thoughts in some of these ways but not in others, and that when one denies that an inserted thought is one’s own, what one is denying is not awareness of the thought but some thicker form of the thinking-relationship.

Adopting this general strategy, Coliva (2002) argues that when the sufferer says ‘X is not my thought’, what he is denying is a sense of authorship of the thought, but not a sense of ownership. All that the sufferer means to convey on this view is that the inserted thought is not integrated with the rest of his cognitive economy (and in this sense, his utterance is correct).

Despite its neat explanation of thought-insertion, however, this account seems to distort what patients and psychiatrists tell us the actual experience of thought-insertion is like. Patients who report thought-insertion do not just say that the thoughts they experience aren’t integrated with their mental economies, they say that these thoughts are not theirs. Patients express feeling alienated from these thoughts in a way that is more substantial than a mere lack of authorship. To appreciate this point, consider the familiar experience of a tune popping into your head. The tune was not prompted by any apparent intention or desire to hear it or anything like it. When this happens, I still recognize this thought as being mine; I will not deny that it is my thought or that it is I who thinks of the tune, even if I ostensibly lack authorship. The schizophrenic agent will, however, call into question whether it is he who thinks the delusional thought. The patients here
aren’t merely saying that their thoughts are authored by other agents, but that their thoughts *belong* to other agents. Indeed, if they were reporting the former then their delusions would belong to an entirely different symptomatic category; they would not be delusions of thought-insertion, but rather delusions of control (DSM-V:87; Frith 1992). By analogy, the experience thus described would not be like my arm being raised by another agent, but rather be like me proprioceptively experiencing an arm that is not mine being raised. I feel alienated from something that I would normally judge to be mine by virtue of the way in which I experience it. So, while innovative, this account fails to adequately capture the apparent phenomenology of delusional experience. This is not to say that the general strategy of interpreting such thoughts as correct is defective, but rather that Coliva’s execution of it is inadequate. There might be a better way to execute this strategy, but because the thinking-relation that is stronger than authorship but weaker than awareness (and which the sufferer must justifiably disavow for his judgment to be correct) is elusive, this strategy is at best incomplete.

Even if this strategy doesn’t work because we can’t lay our hands on the exact thinking-relation needed, there is another, more fundamental, way to slip through the horns of Campbell’s dilemma and to preserve the immunity principle without denying that the thoughts of the schizophrenic are meaningful.

Supposing the patient *is* wrong in thinking (5), he would be wrong not because of misidentification but rather because of mispredication. As we saw in §II, an error of misidentification consists in correctly identifying some property but wrongly identifying the instantiator of that property. What is the property picked out by (5)? Presumably, it is the property of ‘not being the person who thinks ‘KG’’. However, if this is so, then the patient has not correctly identified an instance of this property in the first place; there is no instance of ‘not being the person who thinks KG’ that is correctly identified on the basis of the patient’s experience, for, *ex hypothesi*, he *is* the person who thinks ‘KG’. The immunity principle tells us that any agent who thinks a thought* of the form ‘I am F’ cannot be right on the basis of the perspective grounding this thought* that something is F but wrong about whether ‘I am F’. (5) is not a counterexample to this principle, because, even though he is wrong about whether ‘(5)’ he is not right on the basis of the perspective grounding his thought* that there is someone who is not the person who thinks ‘KG’. Similarly, a delusional patient who thinks ‘I am dead’ would not
be wrong because of an error of misidentification, but rather because his perspective mistakenly picks out an instance of deadness when there is no deadness around. The error could only be one of misidentification if his perspective rightly picks out an instance of deadness but erroneously individuates the bearer of that property. But in each of these cases, the error consists in picking out the wrong predicate (mispredication), rather than in ascribing a predicate that is rightly picked out to the wrong subject (misidentification).

Whether the sufferer is correct or incorrect in thinking (5), the challenge to IP would be unsuccessful: if the sufferer correctly detects an instance of 'person who doesn't think KG', then his claim 'I am not the one who thinks KG' would be error-free. If, however, the sufferer misdetects an instance of 'person who doesn't think KG' then the claim that 'I am not the one who thinks KG' would not be immunized by IP in the first place - the claim would be false because he misdetects a property, not because he misascribes a correctly detected property to the wrong subject.

This section has examined various ways to make sense of Campbell’s claim that thought-insertion challenges our assumptions about the immunity of present-tense introspectively based reports of psychological states to errors of identification. It appears that this claim is wanting. We have seen that it either relies on a tendentious understanding of the scope of the immunity principle or it confuses errors of mispredication with errors of misidentification.

Some might contend that this approach, albeit logically satisfying, has overlooked the more fundamental question of why the patient is delusional in the first place. Although the question of why the cognitive architecture of schizophrenic patients tends to break down in the way that it does is of great significance, it is a project that lies afield of the present analysis, which has aimed to explain the particular question of how the meaningfulness of such patients’ thoughts may be preserved without sacrificing the immunity principle.

V. Integrations and Disintegrations of Selfhood
Nonetheless, this line of critique rightly suggests that while (contrary to Campbell) there is no obvious *logical point* that delusions of thought-insertion seem to contradict, there still seems to be something deeply mysterious about them. Indeed, they represent something radically foreign to ordinary experiences. Our first-person way of knowing is ordinarily bound by an immediately discernible frame of reference fixed by the ‘I’. This frame of reference is given by something akin to what Wittgenstein described as “the inherited background against which one makes judgments about the world…like the axis around which a body rotates” (1969:152). The alignment of this axis and the limits of our experience around it do not permit us to easily understand what it would be like encounter a thought in a first-personal way that is felt as not one’s own.

What is so peculiar about thought-insertion delusions is that someone experiences by way of his distinctive first-person way of knowing what strikes him as thoughts that are not his own. The possibility of such intimate acquaintance with the felt products of another mind challenges a natural assumption about the unity of the self and the reflexive ability to recognize the boundaries between the self and the non-self. The strangeness of the inserted thought thus is not a feature of its content per-se but rather of the disturbed background from which it manifestly issues, viz., an evident breakdown of the self. This breakdown seems to occasion a felt contraction of the parameters of the self that accompanies the sufferer’s distorted grip on reality and failure to recognize his thoughts as his own. By ‘contraction’ of the self, I mean a sense of dwindling away or lack of robustness of one’s own self-boundaries such that one loses the ability to reliably identify with parts or products of one’s own body or behavior. In one person’s account, “I am no longer myself… I feel strange, I am no longer in my body, it is someone else; I sense my body but it is far away, some other place. Here are my legs, my hands, I can also feel my head, but cannot find it again. I hear my voice when I speak, but the voice seems to originate from some other place” (Parnas 2003). What is at once revealing and puzzling about this account and others like it is the use of reflexive possessive pronouns to disavow relatedness to the very events or objects they predicate (e.g., “I am no longer in my body”). Another person with schizophrenia reports, “I seem to be changing and I can’t do anything about it. I feel I am losing myself more each day. That’s bad enough but it’s the vagueness of the whole thing that really troubles me. If the things that were happening were clearer so you could put them into words and
tell somebody what it’s like without sounding quite daft it wouldn’t be so bad” (Flanagan et al. 2010). Similar motifs pervade Elyn Saks’ experiences with schizophrenia: “This experience is much harder, and weirder, to describe than extreme fear or terror. But explaining what I’ve come to call "disorganization" is a different challenge altogether. Consciousness gradually loses its coherence. One’s center gives way. The center cannot hold. The "me" becomes a haze, and the solid center from which one experiences reality breaks up like a bad radio signal. There is no longer a sturdy vantage point from which to look out, take things in, assess what's happening…No organizing principle takes successive moments in time and puts them together in a coherent way from which sense can be made” (Saks 2007). The inability of many patients suffering with schizophrenia to reliably distinguish between internal and external stimuli (e.g., speech that is spoken versus internal dialogue) (Woodward 2013) further illustrates how a profoundly dysregulated sense of self might play out. Ordinarily, individuals have a strong handle on their self-boundaries; distinguishing self from non-self is thus for most individuals an elementary task that can be performed with a high degree of fidelity. As Rhodes & Gipps (2008) describe, this ability relies on a functional bedrock or background framework in relation to which experiences can be made sensible, thoughts can be appropriately contextualized, and in which judgements can be grounded. Instances of thought-insertion provide us with a compelling examples of what ensues when this bedrock is distorted or dissolved; the ability to dependably identify the boundaries of one’s self and the products of one’s own mind apparently degenerates, resulting in mental events that, defying ordinary experience, are ascribed to other agents.

Distortions of perceived self-boundaries underlying delusional experiences could manifest in a variety of ways. Whereas many of the examples considered above result from what I have termed a felt contraction or narrowing of self-boundaries involving a failure to identify with parts or products of ones self, patients may alternatively manifest with felt expansions of self-boundaries wherein one identifies with or experiences felt influence over objects, events or persons beyond ones self. In an uncanny fashion, patients with pathologies of this kind may thus come to identify themselves or their thoughts with people they are not or with inanimate objects in their vicinity, such as tables, chairs or buildings. These delusions reflect profound disturbances in the integrity of the infrastructure that normally confers dependent and well-circumscribed perceptions of where the self begins and ends.
In Schreber’s lucid account of his own mental illness he remarks, “everything that happens is in reference to me … I [am] in a way … the only human being, or simply the human being around whom everything turns, to whom everything that happens must be related” (2000,197). Schreber’s experiences here reveal the reverse of a contraction or dwindling away of the self: what he ostensibly experiences is a phenomenological expansion of the parameters of the self. This uncanny inflation of the perceived limits of the self gives rise to the peculiar impression of being the axis around which everything else turns. These examples show how seemingly dissimilar pathologies, if understood as pathologies of selfhood, may be more alike then we might have thought; approaching and analyzing delusional experiences as pathologies of selfhood, viz., as distortions of the bedrock that regulates the felt parameters of the self, instead of as more circumscribed pathologies of higher order logical or functional faculties may accordingly yield deeper and more fruitful understandings of these conditions and possible approaches to treatment.

As we saw earlier, attempts at demystifying thought-insertion run the serious risk of oversimplifying the content of the delusional experience. I have endeavored here to sketch what I think is more compelling explanation of such phenomena, one in which the scaffolding of the self and the cognate bedrock upon which it is built fall into disrepair, the perceived boundaries of the self are contorted, and uncanny impressions of the world and one’s place in it powerfully ascend through the landscape of the mind.

A version of the analysis advanced in this chapter is forthcoming in Philosophy, Psychiatry & Psychology, Johns Hopkins University Press, 2016.
References


Campbell, J. 1999. Schizophrenia, the space of reasons, and thinking as a motor process. The Monist 82 no.4: 609–625.


Chapter 2:
Mirror Synesthesia and the Limits of Misidentification

[Note: this chapter responds in part to two commentaries forthcoming in Philosophy, Psychiatry & Psychology (Johns Hopkins University Press) by Dr. Clara Humpston and Dr. Lauren Ashwell on my analysis of thought insertion. The full texts of these commentaries are included in Appendices 1a and 1b for readers’ reference.]

In Privileged Access and the Agent in Thought Insertion and Possibilities of Misidentification, Humpston and Ashwell respectively argue for fine-tunings of the approach to thought-insertion advanced in Pathologies of Thought and First Person Authority (hereafter ‘Pathologies’). The present chapter illustrates how the insights that they offer can be understood in the context of the approach developed and defended in Pathologies, and in so doing allow for a more robust understanding of thought-insertion and immunity to error. After critically evaluating key elements of Ashwell’s and Humpston’s arguments, I turn to an exploration of the phenomenon of mirror synesthesia, explaining its meaning and relevance to the issue of immunity to error through misidentification (IEM), and its bearings on theories of selfhood and how to reliably delineate the mechanisms and limits of misidentification.

I. Evidence and Immunity to Error

In Possibilities of Misidentification, Lauren Ashwell argues that IP “doesn’t show us anything about introspection or the first person – which should make us wonder whether it really captures that’s at stake in discussions of IEM.” Ashwell’s argument hinges on two claims:

1. IP turns on features that are not unique to introspection, to the first-person, or to “subject matter that is thought to have IEM,”

2. IP does not yet capture “the standard claim of IEM for introspective self-attribution.”

Ashwell thereupon contends that the immunity furnished by IP “might be merely contingent.” While the structure of the argument in Pathologies may perhaps be used to defend other versions of IEM principles (and whether it does or does not is not obviously consequential for the positions advanced on its basis), Ashwell’s example of BIP – a hypothetical immunity
principle involving Bates College - as a version of an IEM principle that does not involve first-personal introspective judgements is plainly defeasible in a way that IP is not.

To illustrate the shortcomings of BIP, suppose that a first-time visitor was traveling through Lewinston and encountered a particular wall of Bates College with a reflective, mirror-like finish. From this agent’s and only this agent’s perspective, Bates College appears sky-blue by virtue of the reflective angles and his particular position with respect to the light source. On the basis of this impression, this agent forms the thought ‘Bates College is sky-blue.’ While the agent is right on the basis of this evidence that something is sky-blue (viz., the sky), the agent is wrong about whether Bates College is sky-blue. The blueness that the agent identifies is a property of the sky, and not of Bates College. The agent in this setting commits an error of misidentification in a manner that straightforwardly violates BIP. The property of sky-blueness in this context is gleaned from a virtual image of the sky, formed on the basis of the agent’s optical system receiving input from diverging rays reflected from the wall’s plane mirror surface and being processed in a manner that makes it seem as though the rays originate from a real surface behind the mirror when in fact they do not.

Suppose that this agent then walks another few blocks, and hears loud music and a cacophony of voices cheering, laughing and shouting. These sounds seem to the agent to be emanating from a Bates College building, and so the agent thinks on the basis of this evidence that ‘Bates College is a place where people are rowdy right now’. Unbeknownst to this agent, the perceived hubbub is originating from a football field behind Bates College that belongs to a local high school hosting an annual sporting competition. The agent is correct that there is some place where people are rowdy right now, but the agent is wrong about whether Bates College is the place where people are rowdy right now. In this case, the information that makes it sensible for the agent to think that rowdiness is instantiated does not concomitantly individuate the entity that instantiates that property. This point is highlighted if we consider what would transpire if the agent were blindfolded during this first time stroll through Lewinston; the information that would make it sensible to think that rowdiness is instantiated somewhere would not concomitantly individuate what in particular instantiates this property. The epistemic separability that plainly obtains here between the process of property-individuation and the process of property-ascription is distinctively absent in ordinary cases of privileged access. This case and
others like it are thus unlike thoughts* arrived at through the first-person perspective that take as
their subject the essential indexical ‘I’ – those characterized by IP.

Ashwell might counter that in the foregoing cases the evidence at hand is not evidence about Bates College; however, Ashwell does not offer a plausible account of evidence that could support such a contention. Consider how we discuss and evaluate evidence in daily affairs. Suppose for example a patient presents to a neurology clinic with a one-month history of progressively worsening choreiform movements (involuntary arrhythmic, rapid jerky movements of the extremities and face). This pattern of involuntary movements is a hallmark, albeit not pathognomonic, of Huntington’s disease, an autosomal dominant genetic neurodegenerative disorder. The neurologist on service states to a medical student on the team “this patient’s choreiform movements are evidence of Huntington’s disease.” If it turns out that the patient’s choreiform movements are in fact not due to Huntington’s disease but rather due to a rarer autoimmune etiology (cf., O’Toole 2013), the neurologist’s former evidentiary statement is by no means made false; one may still rightly cite the symptom as evidence in favor of a diagnosis of Huntington’s disease, even if all-things-considered this diagnosis would be misplaced. Similarly, if the patient received a blood test and another patient’s results with the same name were mistakenly printed by the laboratory and sent to the clinician, and the results suggested a leukocytosis, the clinician may still justifiably think ‘these laboratory results are evidence that my patient is leukocytotic’. Evidence in this context, as it seems to operate in most contexts, suggests but does not indefeasibly demonstrate a state of affairs that was previously ambiguous or unknown.

The epistemic gap between suggestion and indefeasible demonstration with respect to evidence opens up the possibility for an agent to justifiably take as x as evidence for p in situation z even if it is the case that ~p in z. Hence in ordinary language we have terms such as ‘misleading evidence’, ‘unreliable evidence’, ‘ambiguous evidence’, etc. While a more lengthy explication of the epistemological and logical dynamics of this three-place relation and the modulatory role of the agent’s perspectives therein are beyond the scope of the present analysis, the foregoing discussion reveals that the agent in the cases above can justifiably be thought to have evidence about Bates College.

Evidence could thus mislead. As J.L. Austin incisively describes in Sense and Sensibilia, “[t]he situation in which I would properly be said to have evidence for the statement that some
animal is a pig is that, for example, in which the beast itself is not actually on view, but I can see plenty of pig-like marks on the ground outside its retreat. If I find a few buckets of pig-food, that's a bit more evidence, and the noises and the smell may provide better evidence still. But if the animal then emerges and stands there plainly in view, there is no longer any question of collecting evidence; its coming into view doesn't provide me with more evidence that it's a pig, I can now just see that it is. The question is settled” (Austin 1963). If the pig-like marks turn out to be marks of a different animal’s tracks, or the handiwork of an artist specializing in etching animal tracks into the earth, or a holograph projected from far away, then the marks do not then fail to be evidence suggesting that there is a pig around, despite the fact that they mislead us about whether there is in fact a pig around. These general features of evidence suggest that the relationships between pieces of evidence and the conclusions they support are sufficiently elastic to withstand settings in which evidence points to conclusions that do not align with actual states of affairs, or claim as subjects entities upon which individuated properties are misattributed.

Most notably, regardless of the approach one takes to evidence, IP, unlike BIP, does not explicitly hinge on evidence but rather on grounding perspectives that may function through mental activity that is not evidence-like in nature in the first place. In first-personal cases, as has been argued previously, the information that makes it sensible for one to think that there is some property being instantiated concomitantly individuates the thing that instantiates the property. Arriving at thoughts about oneself is in this way unlike thoughts typically arrived at about other people or other things.

Ashwell curiously asserts that “it is only contingent that other subject matters lack such an exclusive “way” to obtain grounds for judgments about them; we can surely imagine having a faculty whose exclusive use is for coming to judgments about Bates College,” yet Ashwell concedes that “such a possibility is definitely a strange one” (p. 5). It is difficult to imagine what it would be like to have such intimate acquaintance with someone or something other than oneself. Notwithstanding, even if robustly conceivable, it would not follow from such possibilities that IP is not grounded in something peculiar to privileged access, the first-personal or introspection, as Ashwell controversially submits. Indeed, such possibilities might hinge on aberrant amplifications of typical introspective capacities, expansions of the perceived boundaries of selfhood, or distortions of self-other processing to enable felt intimate acquaintance with beings or objects that are beyond the typical limits of routine first-person
experience. This possibility is returned to in the later discussion of the phenomenon of mirror
synesthesia. Importantly, the conceivability of introspection reliably delivering information
about something other than one’s own mental states, while deeply interesting, does not pose any
obvious logical challenge to IP.

II. Privileged Access and the Immunity Principle

A related point is raised by Clara Humpston in *Possibilities of Misidentifitication*, where it
is argued that a more appropriate term than ‘delusions of thought insertion’ may be ‘delusions in
thought insertion’, since it appears that the experience of an external thought through the first-
person perspective alone is insufficient for a delusional elaboration to assemble. Instead,
Humpston contends that a separate act of ascribing another agent to such a thought is required to
reach a delusional end-point (p. 4). Two key points are embedded in this thesis, one concerning
the non-delusional character of thought-insertion simpliciter, and the other concerning the
process by which thought-insertion might attain a delusional character.

Although Humpston predicates this theory on an account that bisects the delusion into an
inserted-thought and a false belief about the inserted-thought, it is not obvious that a doxastic
conception of delusions is required for either of these two points. A more plausible conception of
thought-insertion delusions may well be non-doxastic, especially given that such delusions do
not appear to reasons-responsive or doxastically integrated as normal belief-states are.

In their provocative piece *Delusion, Rationality and Empathy* (2001), Currie and
Jureidini compellingly argue along these lines for the position that delusions hinge not on beliefs
but rather on imaginings:

“There are occasions on which all of us form beliefs immediately on the basis of
experience but reject them quickly thereafter when we find evidence that they are wrong. That is
how it is with someone who forms the belief that one line is longer than another in a visual
illusion experiment. Why do deluded subjects not reject their delusions when they see that they
are inconsistent with so much else that they do accept? The hypothesis that delusions are
imaginings explains this rather well. When someone acquires a new belief, a potentially
wholesale process of belief revision will take place; the person ceases to believe things she
previously believed, because they are inconsistent with the new belief. Failures to resolve
clashes between older beliefs and new beliefs will, if made evident to the subject, seem
problematic to her. However, when we imagine something, we do not cease to believe things that are inconsistent with what we imagine, and we do not feel any pressure to resolve clashes between what we believe and what we imagine. Rather, beliefs inconsistent with the imagining move temporarily into the background; they are not available as premises in inferences that involve the imagining but are available as premises in inferences that do not. Finally, imaginings are not apt to be revised in the light of evidence; the whole point of imagining is to enable us to engage with scenarios that we know to be nonactual. Thus, imaginings seem just the right things to play the role of delusional thoughts; it is of their natures to coexist with the beliefs they contradict, to leave their possessors unwilling to resolve the inconsistency, and to be immune to conventional appeals to reason and evidence... What is emerging here is a picture of (some) delusions as states which, on the one hand, are imaginings but on the other, have a feature that most imaginings lack: a feeling of subjective conviction on the part of the subject... The subject who can no longer tell an imagining from a belief may yet retain the capacity to see that some particular delusional item is wildly inconsistent with the rest of her epistemic corpus” (160-161).

If Currie and Jureidini are correct, then even if a non-doxastic account of delusions is maintained, Humpston’s point regarding the necessity of a separate act of ascribing another agent to inserted thoughts to reach delusional end-points might still be upheld if such ascriptions proceed through imaginative processes rather than through a doxastic processes (though the term ‘act’ here would likely require further explication and possible revision.)

Regardless of the conceptual viability of separating the experience of TI from delusions of TI, Humpston’s later claim that delusions in TI are susceptible to IP-relevant errors of misidentification does not follow from the strategy. As detailed in *Pathologies*, the mere ascription of the wrong instantiator to a thought that is in one’s mental space does not constitute an IP-relevant error of misidentification if the thought does not take as its subject the indexical ‘I’. Moreover, when thoughts* do take as their subject the essential indexical, the psychological processes of property-individuation and property-instantiation become increasingly inseparable. Humpston’s hypothesis that “the patient is not delusional at first when they experience the ‘inserted’ thoughts but becomes so when they arrive at the conclusion that they are thought by an external thinker," albeit highly interesting, is thus not in any obvious sense incompatible with IP.
Reflecting on how these phenomena bear on notions of selfhood, Humpston points out that a thinker may misattribute thoughts that threaten the thinker’s subjectivity (pp. 6-7). Though Humpston suggests that this might generate a paradox given that this mechanism for protecting subjectivity itself necessitates some putative prior fragmentation of the self, there is no paradox here; the misattribution of thoughts should be understood as a symptom, and not as a source, of fragmented selfhood or unstable subjectivity. This view is deeply in tune with the account of selfhood defended in Pathologies and explore further below.

The notion that delusions may play adaptive roles in patients’ lives is compelling. As Hundert intriguingly explores in The Brain's Capacity to Form Delusions as an Evolutionary Strategy for Survival (1992) the neurobiological paradigm often tendentiously attempts to isolate the locus of delusions in a 'broken brain', and clinicians’ inability to definitively 'localize the lesion’ is often mistaken as a sign of neurologic complexity rather than as a sign that the approach itself may be misguided (349). Hundert presents the fascinating case of Timothy G., a young man prone to episodes of severe depression and suicidal ideation who develops a delusional system wherein whenever he goes through a severe depressive episode he experiences that he is Hitler reincarnated to do penance for his war crimes. This case elegantly demonstrates how psychiatric sympamotology and phenomenology might play a pivotal role in individual adaptation and in the maintenance of subjective integrity; through Timothy G’s delusional system, “meaning was returned to his life...[for] to end his life would be to end the world’s only hope for bringing justice to the perpetrator...[and] since the time his brain figured this out, his continued survival has not been in jeopardy” (351). The emergence of this delusional alter thus gave form to an adaptive and stabilizing weltanschauung in the midst of an otherwise unbearable and functionally disintegrating lifeworld. This notion squares well with Eugène Minkowski’s observation that “a delusion is not completely a product of the imagination. It becomes grafted onto a phenomenon which is a part of our life and comes into play when the life-synthesis begins to weaken....[and it] is in fact only the attempt of that particular part of the mind which remains intact to establish a logical connection between various sections of a crumbling edifice” (Hundert 1992: 352; Minkowski 1970:188). The delusional system at play in many cases of TI may thus represent an attempt to maintain subjective integrity in the face of a crumbling edifice of the self.
III. Mirror Synesthesia and the Malleable Self

Both Humpston’s hypothesis regarding the necessity of a separate act of ascribing another agent to an inserted thought to reach a delusional end-point and Ashwell’s point that introspection can reliably deliver information about another gesture intriguingly toward possibilities for non-delusional felt first-personal-like awareness of mental states that are not one’s own.

The condition of mirror synesthesia delivers empirical support for such possibilities and provides a novel and illuminating lens through which the mechanisms and limits of privileged access and (mis)identification might be better understood. The phenomenology and neuroscience of mirror synesthesia not only demonstrate that first-personal-like awareness of another’s mental states may be possible, but also blur the boundaries between identification and misidentification, and challenge philosophical orthodoxy about the self in support of a dynamic and ecological scaffolding of selfhood.

Mirror synesthesia is a condition wherein subjects experience the sensations or emotions of another observed individual. For instance, observing another’s hand being stroked would evoke in the mirror synesthete a concomitant tactile sensation of the synesthete’s own hand being stroked. Originally dubbed mirror-touch synesthesia, terms such as mirror-sensory synesthesia or mirror-sensory experience are now preferred by many to capture the fact that synesthetes’ mirror experiences may span beyond touch to include pain, emotions and other apparent mental states (Keysers & Gazzola 2009). Bradshaw and Mattingley report the illustrative case of a man whose wife observes “[i]f I slightly knocked my finger, spontaneously showing him, he would immediately grasp his own finger and say “don’t do that” (meaning not to show him); He actually felt it. If I merely commented (that I had knocked my finger), there was no sensation” (2001:135). Some oppose use of the term ‘synesthesia’ altogether for this condition since it does not exhibit crossmodal sensory activity but instead exhibits intramodal activity contained to the same type of sensory domain (Rothen & Meier 2013). Here I simply employ the term ‘mirror synesthesia’ to remain inclusive to mirror experiences that are not of sensations per-se but of other observed states as well.
Emerging empirical evidence suggests that mirror synesthesia may result from atypically enhanced cortical excitability within brain regions responsible for normal somatosensory mirroring and self-other processing, regions which have been implicated in the contagious character of phenomena such as laughter and yawning (Bannissy and Ward 2013:3; Northoff 2011; Brent 2014). However, there exists some ambiguity and perhaps confusion within the neuroscientific literature surrounding what it is that mirror synesthetes precisely experience when undergoing mirror-experiences. This may be in part due to the existence of multiple subtypes of mirror-synesthesia, each which might exhibit a potentially different phenomenology. For example, whereas some might experience an observed state as being superimposed or duplicated in their own body, others might experience the observed state as if they actually co-inhabited the other body. Still others might describe a loosening of the ability to distinguish self and other observed bodies, occasioning what is felt to be a token-identical in contrast to a type-identical experience. Depending on the phenomenology, different descriptions of mirrored experience might therefore ensue.

Many judgments formed on the basis of mirror synesthesia appear to withstand the test of IP. For example, a mirror synesthete who expresses ‘I am in pain’ when observing another in

---

4 The evidence for normal mirroring is described in Keysers and Gazzola (2009): “Whenever we see what happens to others, we not only understand what they experience but [we] also empathically share their states. In the nineties a series of experiments carried out in Parma evidenced that some premotor neurons, called mirror neurons, fired not only during action execution, as a premotor neuron ‘should’, but also during the observation of the same action… we [also] might empathically share the states of others because seeing their states triggers representations of corresponding states in our brain… mirror neurons [which serve as the neural substrate for such processes, and respond] during the execution and perception of actions not only exist in the premotor cortex… but also other brain regions and species [with] similar properties. Second, we also recruit brain regions responsible for our own sensations and emotions while we perceive those of others, and these various systems are intimately interconnected.. these new findings encourage us to see mirror neurons in the premotor cortex as the best understood example of a wider class of vicarious neural responses encompass… there is now hard evidence that at least five brain regions of the human cortex contain mirror neurons: the ventral and dorsal premotor cortex, the supplementary motor cortex, the inferior parietal lobe, and the temporal lobe.” 666-667) Interestingly, these areas are largely coextensive with those correlated with the regulation of self-concepts and self-awareness. These findings suggest that mirror sensory experiences may hinge on alterations in systems maintaining self-concepts and regulating self-other representations. Such alterations in turn may give rise to enhanced activity in mirror-neuron networks. Empirical support for this approach can be found in studies that demonstrate that the degree of activity within mirror-neuron systems during observation of another tends to track the degree to which the subject judges the other to be similar to the subject’s self (Santiestaban et al. 2015; Serino et al. 2009).

5 Other descriptions that might be produced by mirror synasthetes, such as “I feel what he is feeling” are notably ambiguous between these different potential phenomenologies.
pain commits no error of misidentification, for they are not right that someone is in pain but wrong about who is in pain; the mental state of pain is in fact experienced by the synesthete, albeit occasioned by the external stimulus of the observed individual in pain. Further conceptual and empirical study will be needed however to reveal whether errors of misidentification may occur in extreme cases in which synesthetes feel as though they are the ones undergoing the token experiences of the observed individual (if such cases do in fact exist). Imagine for instance a mirror-synesthete observing another’s hand being stroked by the paw of a dog. If on the basis of this experience the synesthete sincerely expresses ‘I am someone whose hand is being stroked by a dog,’ the philosophical implications would be quite different from the alternative, ‘I am someone who feels that my hand is being stroked by the dog’.

Some may argue that the former possibility challenges IP. However, such arguments are significantly confounded by emerging conceptions of the self as a malleable construct—conceptions with significant implications for how the essential indexical ‘I’ ought to be conceptualized and operationalized. These notions are turned to in greater detail in Chapter 3 - Constructing the Essential Indexical.

Intriguingly, recent work by Maister et al. (2013) suggests that mirror synesthetes undergo measurable changes in self-representations when observing others’ sensory experiences. When mirror-synesthetes took part in an enfacement illusion experiment in which they observed unfamiliar faces being touched and were later queried about their experiences, sentiments such as “I felt like the other’s face was my face” and “I felt like my own face began to resemble the other person’s face” (2013:804) were widely endorsed, demonstrating intriguing malleability of self-other boundaries by way of what the investigators characterize as an “interjection of the other into the self…a change in the mental representation of the self” (2013:807). In a more recent study, Banakou and Slater (2014) demonstrate that subjects who experience a life-size virtual body (VB) simulation seen from the first-person perspective have a “strong subjective illusion of body ownership and agency with respect to the VB” and misattribute vocalizations of the VB to themselves (endorsing the statement “It felt as if I was speaking out the words I heard”) after experiencing a brief experimental condition of synchrony between the VB and the real body. Subjects also “shifted the fundamental frequency of their later utterances toward the stimulus voice” (2014:17678)
In addition to providing intriguing examples of non-delusional felt first-personal awareness of mental states that are not one’s own, instances of mirror synesthesia thus deliver further evidence for the dynamic and interrelational nature of the self, a self that forms and reforms with ongoing processing of internal and external stimuli, a self that expands and contracts in a manner not bound by the perimeter of the body.

The elasticity of selfhood suggested by these findings allows for a novel conceptualization of self-other representation as a broad and continuous spectrum wherein phenomena such as mirror synesthesia lie on one end and phenomena such as thought-insertion delusions in schizophrenia lie on the other. Whereas mirror synesthesia manifests with a blurring of the self-other boundary in the outward direction (thereby permitting first-person-like identification with another body’s sensations or emotions) TI in schizophrenia may manifest with a blurring of the self-other boundary that is pathologically inward (thereby potentiating failure to recognize thoughts that emerge in one’s own mind as one’s own). Intriguingly, the directionality of blurring of the self-other boundary could be constitutive of processes governing empathic capacities; whereas mirror synesthesia is linked to significantly heightened empathic abilities (Bannissy & Ward 2007), schizophrenia is linked to significantly deficient empathic abilities (Derntl et al 2009).

The potential for alterations in self-representation to modulate empathy along a continuous axis is one of many potential examples of how understanding self-representation as a spectral function can allow for better characterization of similarities and differences between various pathologies of the mind; of their correlates in thought, emotion and behavior; and of the phenomenological, neurobiological and philosophical landscapes that converge and diverge in ways endlessly profound.

A distilled version of this chapter’s analysis, along with the complete commentaries to which it responds, are forthcoming in Philosophy, Psychiatry & Psychology, Johns Hopkins University Press, 2016.
References


Ashwell, Lauren. 2015. Possibilities of Misidentification.


Chapter 3:
Constructing the Essential Indexical ‘I’

"A main cause of philosophical disease - one-sided diet: one nourishes one's thinking with only one kind of example" – Ludwig Wittgenstein, PI §593

“One of the most mysterious of semi-speculations is, one would suppose, that of one Mind’s imaging into another” - John Keats, margin note in his copy of Paradise Lost, 4:1.59-94

As considered in the preceding chapters, the phenomena of thought-insertion and mirror synesthesia provide instructive pathways through which to probe the nature of selfhood. In particular, as we have seen, careful examination of these phenomena bring to the fore the notion of self-other representation, revealing at once its role in the elaboration of individual identities while also revealing its remarkable porosity and elasticity. Porosity, insofar as the territory mapped out by the cognitive boundary assembled between self and other appears to be constitutively defeasible (temporally or substructurally), and elasticity, insofar as this boundary is not static but appears to dynamically expand or contract.

Inter alia, the configuration and function of this capacity in any particular instance will influence which entities in physical and psychological space an individual identifies with (viz., relates to as their own) or fails to identify with (viz., relates to as foreign). There is far more to selfhood, however, than is dreamt of in this philosophy. What is constitutive of selfhood beyond this capacity, and the relevance of this issue to medical practice and research, is the focus of the present chapter.

I. Selfhood Beyond Self-Other Representation

In the latter half of his Philosophical Investigations, Wittgenstein reminds us of “a main cause of philosophical disease - one-sided diet: one nourishes one's thinking with only one kind of example." In considering the constitution of selfhood, it is important to relinquish this impulse, and to seek conceptual nourishment from as wide array of examples as possible.

Consider an individual who wholly lacks the intact capacity to distinguish between self from other; in such a scenario, selfhood does not altogether vanish in the absence of this intact capacity. A self remains, albeit noticeably less robust. It is neither a shadow of selfhood nor a
mimic of selfhood that one identifies in observing such an individual, rather, it is a self that is in one, rather particular sense, indisposed. The limited extent of the explanatory power of this capacity alone in accounting for what is constitutive of selfhood will be more fully appreciated in the course of accounting for other key components of this complex scaffolding, and in directing attention to a range of clinical examples and thought experiments not ordinarily considered.

To illustrate the nature of traditional problems that arise in the course of theorizing about selfhood and key approaches taken toward addressing them, however, a few central points regarding the notion of selfhood in the history of modern philosophy are first in order.

II. Beyond Hume: Demystifying the Scaffolding of Selfhood

In *A Treatise of Human Nature*, Hume famously casts skepticism on the foundations of philosophical reflection on the self, arguing that since the self is never directly observed, and rather what is observed is only successions of experiences (never the underlying putative self having those experiences), to speak of the self is empirically untenable (my emphasis):

“There are some philosophers who imagine we are every moment intimately conscious of what we call our self; that we feel its existence and its continuance in existence; and are certain, beyond the evidence of a demonstration, both of its perfect identity and simplicity…but self or person is not any one impression, but that to which our several impressions and ideas are suppos'd to have a reference. If any impression gives rise to the idea of self, that impression must continue invariably the same, thro' the whole course of our lives; since self is suppos'd to exist after that manner. But there is no impression constant and invariable. Pain and pleasure, grief and joy, passions and sensations succeed each other, and never all exist at the same time. It cannot, therefore, be from any of these impressions, or from any other, that the idea of self is deriv'd; and consequently there is no such idea…[what is experienced is] nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement. Our eyes cannot turn in their sockets without varying our perceptions. Our thought is still more variable than our sight; and all our other senses and faculties contribute to this change; nor is there any single power of the soul, which remains unalterably the same, perhaps for one moment. The mind is a
kind of theatre, where several perceptions successively make their appearance; pass, re-pass, glide away, and mingle in an infinite variety of postures and situations. There is properly no simplicity in it at one time, nor identity in different; whatever natural propulsion we may have to imagine that simplicity and identity. The comparison of the theatre must not mislead us. They are the successive perceptions only, that constitute the mind; nor have we the most distant notion of the place, where these scenes are represented, or of the materials, of which it is compos'd” (Hume 533-534).

This variety of empiricism, on which the meaningfulness of a concept is necessarily grounded in direct experiences of the concept, leads Hume not only to reject the meaningfulness of the self but also to reject the meaningfulness of the concept of causation; while sequences of events may be observed, inability to directly observe the causal glue, as it were, that links events in a putatively causal scheme, animates Hume’s skepticism about causation in a manner similar to his skepticism about selfhood.

Responses to Humean skepticism about selfhood have classically assumed one of two forms, one rooted in the tradition of René Descartes and one rooted in the tradition of Immanuel Kant.

In the view of Descartes, the mere existence of mental activity analytically entails a thinking self. Accordingly, he argues in Meditations on First Philosophy, “[b]ut what then am I? A thing which thinks. What is a thing which thinks? It is a thing which doubts, understands, affirms, denies, wills, refuses, which also imagines and feels... Here I make my discovery: thought exists; it alone cannot be separated from me. I am; I exist – this is certain. But for how long? For as long as I am thinking; for perhaps it could also come to pass that if I were to cease all thinking I would then utterly cease to exist ...I am therefore precisely nothing but a thinking thing; that is a mind, or intellect, or understanding, or reason – words of whose meanings I was previously ignorant. Yet I am a true thing and am truly existing; but what kind of thing? I have said it already: a thinking thing...it is this same "I" who senses or who is cognizant of bodily things as if through the senses. For example, I now see a light, I hear a noise, feel heat. These things are false, since I am asleep. Yet I certainly do seem to see, hear, and feel warmth. This cannot be false. Properly speaking, this is what in me is called "sensing". But this, precisely taken, is nothing other than thinking” (31-32, my emphasis).
An important upshot of this argument is that it results in selfhood being bound up with consciousness. In Descartes’ view, then, a thing that does not think cannot be justifiably called a self, and any thinking thing might be said to instantiate some degree of selfhood.\(^6\) In the language of contemporary neuroscience, which canonically describes consciousness as the union of wakefulness (i.e., arousal, which is mediated by the reticular activating system and its ascending thalamocortical projections traversing the rostral brainstem tegmentum) and awareness (mediated by the cerebral cortices, thalami, basal ganglia, and limbic networks), it is presence or absence of awareness that is constitutive of the ‘I’. (Interestingly, such considerations intersect with those used in prevailing approaches to distinguishing brain death from vegetative states, with brain death hinging on irrecoverable loss of wakefulness and awareness (Cf., Fins 2015; Wijdicks 2015)).

In contrast, the Kantian response to Humean skepticism about selfhood focuses not on the logical entailments of consciousness but rather on a subtle yet salient feature of our experiences themselves, namely, the experience of sequential events as a connected, coherent and synthesized wholes, rather than as disjointed series of discrete occurrences. It is this synthetic unity of experience (what Kant dubs the “transcendental unity of apperception”) that necessitates positing an underlying self that enables the phenomenological unity of experienced sequences of events that would otherwise be experienced as disjointed.

Kant thus zigs where Hume zags. While conceding the Humean point that the self is never empirically observed, rather than concluding that the self does not exist Kant contends that the self is necessarily given a priori as a condition for the possibility of a unified and coherent phenomenology of experience.\(^7\) It is important to note that this synthetic capacity operates in

\(^6\) That selfhood is fundamentally nonmonotonic and may come in go in degrees is a key point of the present chapter.

\(^7\) While a thorough Kantian exegesis is beyond the is scope of the present analysis, this is I believe the thrust of Kant’s argument in the Critique of Pure Reason, where he enigmatically writes that “[c]onsciousness of self according to the determinations of our state in inner perception is merely empirical, and always changing. No fixed and abiding self can present itself in this flux of inner appearances...What has necessarily to be represented as numerically identical cannot be thought as such through empirical data. To render such a transcendental presupposition valid, there must be a condition which precedes all experience, and which makes experience itself possible. There can be in us no modes of knowledge, no connection or unity of one mode of knowledge with another, without that unity of consciousness which precedes all data of intuitions, and by relation to which representation of objects is alone possible. This pure original unchangeable consciousness I shall name transcendental apperception. That it deserves this name is clear from the fact that even the purest objective unity, namely, that of the a priori concepts (space and time), is only possible through relation of the intuitions to such unity of
consciousness. The numerical unity of this apperception is thus the a priori ground of all concepts, just as
the manifoldness of space and time is the a priori ground of the intuitions of sensibility. This
transcendental unity of apperception forms out of all possible appearances, which can stand alongside
one another in one experience, a connection of all these representations according to laws. For this unity
of consciousness would be impossible if the mind in knowledge of the manifold could not become
conscious of the identity of function whereby it synthetically combines it in one knowledge. The original
and necessary consciousness of the identity of the self is thus at the same time a consciousness of an
equally necessary unity of the synthesis of all appearances according to concepts, that is, according to
rules, which not only make them necessarily reproducible but also in so doing determine an object for
their intuition, that is, the concept of something wherein they are necessarily interconnected...The
consciousness of self is thus very far from being a knowledge of the self, notwithstanding all the
categories which [are being employed to] constitute the thought of an object in general, through
combination of the manifold in one apperception. Just as for knowledge of an object distinct from me I
require, besides the thought of an object in general (in the category), an intuition by which I determine
that general concept, so for knowledge of myself I require, besides the consciousness, that is, besides the
thought of myself, an intuition of the manifold in me, by which I determine this thought. I exist as an
intelligence which is conscious solely of its power of combination; but in respect of the manifold which it
has to combine I am subjected to a limiting condition (entitled inner sense), namely, that this combination
can be made intuitable only according to relations of time, which lie entirely outside the concepts of
understanding, strictly regarded. Such an intelligence, therefore, can know itself only as it appears to
itself in respect of an intuition which is not intellectual and cannot be given by the understanding itself,
not as it would know itself if its intuition were intellectual. The 'I think' expresses the act of determining
my existence. Existence is already given thereby, but the mode in which I am to determine this existence,
that is, the manifold belonging to it, is not thereby given. In order that it be given, self-intuition is
required; and such intuition is conditioned by a given a priori form, namely, time, which is sensible and
belongs to the receptivity of the determinable [in me]. Now since I do not have another self-intuition
which gives the determining in me (I am conscious only of the spontaneity of it) prior to the act of
determination, as time does in the case of the determinable, I cannot determine my existence as that of a
self-active being; all that I can do is to represent to myself the spontaneity of my thought, that is, of the
determination; and my existence is still only determinable sensibly, that is, as the existence of an
appearance. But it is owing to this spontaneity that I entitle myself an intelligence.” Consider the
following illustration of Kant’s idea. In order to experience a piece of music as a concerto, the
transcendental unity of apperception must be constantly active, there must be some self-aware “I” which
grounds experience altogether: if this were not the case, no piece of music could ever be experienced as
such, instead, one would simply hear one note after another without any conception of linked unity or
harmony between them in the aggregate. There must in his view be some subject which can bear the self-
conscious experience of music. It is in this way that “no cognitions can occur in us, no connection and
unity among them, without that unity of consciousness that precedes all data of the intuitions...this pure,
original, unchanging consciousness ...[is named the] transcendental apperception” (Kant CPR 232).

The similarity to space and time that Kant draws here is apposite. Kant argues for the position
that, like the self, space and time are a priori forms of pure intuition, which are superimposed by the
subjective mind onto sensibility and thereby frame perception of the external world. This is in sharp
contrast to a Newtonian view, which would proffer that space and time are external objects, as well as to a
Liebnizian view, which would proffer that space and time emerge from relations between objects. Albeit
differing in their accounts of how independent space and time are from regular objects, both Newton’s
and Liebniz’s views suggest that space and time are external entities which are only later perceived by
minds; on these views, space and time are known only a posteriori. That is, a given mind can acquire the
notions of space and time only following certain experience(s). Kant, however, is proposing the idea that
space and time are not external, but rather a priori, that is, known prior to experience. Kant’s incisively
expresses this view in positing that “there are two forms of sensible intuitions as principles of a priori
both diachronic and in synchronic modes (Cf., Brook 1994); diachronic, in its ability to synthesize temporally distinct experiences; and synchronic, in its ability to synthesize numerically distinct experiences.

To illustrate, consider an individual listening to a piece of music while simultaneously playing chess. Such an individual is able to do engage in both of these experiences at once; it is not that she oscillates back and forth between playing chess and listening to music, rather, the capacity for synchronic unity enables one to simultaneously experience both of these activities. On the other hand, the capacity for diachronic unity grounds the ability to experience the piece of music as a coherent whole rather than as a series of disjointed notes, as well as the ability to play a coherent round of chess experiences as a unified series of moves in a game, rather than as a disunified string of unrelated movements of small physical objects of varying shapes. These two crucial capacities served by the self – diachronic synthesis and synchronic synthesis – are important to consider in appraising the varieties of pathologies of selfhood that may emerge, as is explored in further detail in section III.

cognition, namely space and time” (Kant, CPU 174). It is important to note that Kant does not call space and time concepts, but rather forms of intuition. Kant clarifies this point in one of his lectures on metaphysics; space and time, says Kant, “can be only intuition, and never concepts: for to this belongs essentially the immediate representation on objects, and indeed of a single object, and a feature of a concept [which is fundamentally mediate] can never underlie it” (Metaphysik Vigilantius (K3) 446, AK. 29: 975). Metaphysik Vigilantius (K3), notes taken at one of Kant’s metaphysics lectures on Ontology, sheds light on Kant’s understanding of the difference between concepts and intuitions. “Intuition” notes Kant’s student, is an “immediate representation of an object. This...can only be singular. E.g., one sun. I.e., it contains only one object; for were several objects present at once then they could be represented only together through a feature, i.e., be through through concepts.” A concept, however, is the “mediate… exhibition of an object by virtue of a feature common to several objects as a ground of cognition...and thereby the cognition, as a feature, belongs merely to thinking of the object.” Thus, “in intuition the object can be represented insofar as it is given, but through concept the object is representable [only] insofar as it can be thought as given mediately”. Vigilantius notes a final formulation of this distinction: “intuition is sensible, but concepts belong merely to the thinking understanding, and sensibility can be mixed in with them only insofar as they rest on intuition, otherwise...they can have no sensibility through which they are perceived.” (Metaphysik Vigilantius (K3) 442, AK. 29: 971). A proper understanding of this view reveals that for Kant, space and time are transcendently ideal (i.e. made possible solely by virtue of, and understood properly as, not external entities, but rather a priori contributions to sensibility “and cannot be counted as either subsisting or inhering in the objects in themselves without their relation to our intuition) yet metaphysically real (that is, we can still “assert the empirical reality of space [and time]... as a subjective representation related to something external that can be called a priori objective”). Kant reiterates his thesis numerous times and in different ways, but his basic thesis is that “space [and time] represent no property at all of any things in themselves nor any relation of them to each other” but rather they are “nothing other than...the subjective condition of sensibility... [and are] thus a priori” (Kant, CPU, 176-177, 164).
Our analysis thus far has brought to light at least three conceptualizations of what is constitutive of selfhood: (A) that which consists in synthesizing experiences (diachronically and synchronically); (B) that which consists in conscious awareness/thinking; and (C) that which consists in differentiating the personal from the foreign (i.e., self-other representation, which may operate doxastically and affectively). In considering this conceptual landscape, it is most suitable to regard this cast of capacities as symbiotic elements of selfhood, each illuminated by a different sort of example, rather than as logically competing or mutually exclusive theories. The remainder of this section will explore three other capacities to be accounted for in a satisfying scaffolding of selfhood: narrative, ecological embeddedness, and cognitive unity.

**Selfhood as Narrative**

In addition to the capacities of individuals to experience numerically and temporally distinct events as unified, the ideal constitution of selfhood entails a capacity of individuals to weave those events into a coherent and characteristic life-narrative. This capacity may distinctively modulate the meanings, motivations, and attitudes individuals display in reaction to and in anticipation of happenings in their lives.

Most recently, Schechtman (2015) champions this narrative view and offers an illustrative example of its import:

“take an event described neutrally as stepping off of an airplane at a particular airport. For the person leaving the plane, this event will be very different if she is (a) arriving at the country where she will finally pick up the child whose adoption she has eagerly anticipated, (b) arriving at the country of her deployment for dangerous military duty having left small children behind at home, (c) arriving at the place where she will give an academic address and receive a prestigious reward, or (d) arriving for the funeral of a beloved relative who died unexpectedly and very young. The same neutrally described event will yield a different experience and lead to different actions and decisions in each of these cases because it is part of a different narrative and it takes its character from the story of which it is a part. The Narrative Self- Constitution View claims that all of our experiences are like this. As persons interact with the world, they carry with them an implicit awareness of the basic elements of their histories and
anticipated trajectories which at each moment influences both their experience of the present and their deliberations about what to do next. According to this account, *persons are able to have the kinds of experiences and engage in the kinds of behaviors they do precisely because they bring their ongoing life stories to bear on the present and so structure their experience of the world according to an ongoing autobiographical narrative*. The unity of a single person, according to this view, is the unity of a narrative” (2015:396-397, my emphasis).

While Schechtman’s articulation of the capacity to “present and structure the world according to an ongoing autobiographical narrative” is apposite, the position appears to overreach insofar as it asserts that narrative alone is constitutive of selfhood. That the capacity for narrative is not wholly constitutive of selfhood becomes evident in examining a range of clinical cases. Consider, for example, patients with amnestic syndromes arising for instance from lesions to the thalamus, hippocampus, or fornix (Harriot 2015), in some forms of epilepsy (Campora 2016), in forms of dementia, or in autoimmune conditions such as limbic encephalitis (Frisch 2013). In such cases patients may, to varying degrees, lose their abilities to recount or produce coherent narratives of events in their current or past lives, and yet it would be tendentious to posit that selfhood in such patients is perforce lost. Remaining dimensions of selfhood may remain intact despite the loss of narrative unity. Without narrative unity, it is not that selfhood is lost, but rather that selfhood has changed; it becomes, in an important sense, less robust, but it does not altogether recede.

It is important to note that the capacity for constructing active narratives depends, in turn, on a variety of other important capacities, including memory, assignment of meaning and emotional valence, and, as will be explored in the following section, embeddedness in a field of dynamic relationships.

*From Ecological Embeddedness to Selfhood*

Drawing on their anthropological study of self-construal across different cultures as seen through child-rearing behaviors and other social norms, Markus and Kitayama (1994) argue that the concept of the self as an independent unit distinct from the collective is misleadingly overemphasized in European-American theory. Instead, they develop and defend a concept of selfhood as intimately bound up in the greater collective: “[T]he self [should be] viewed not as
an independent entity separate from the collective,” they contend, “but instead as a priori fundamentally interdepending with others. Individuals do not stand in opposition to the confines and constraints of the external collective, nor do they voluntarily choose to become parts of this external collective. Instead the self is inherently social - an integral part of the collective. This interdepending view grants primacy to the relationship between self and others. The self derives only from the individual’s relationships with specific others in the collective. There is no self without the collective; the self is a part that becomes whole only in interaction with others... in this view a type of intersubjectivity, rather than a private subjectivity, would be the strongest, most elaborated aspect of self” (1994:570).

As an example that illustrates how the self may achieve elaboration only within a social collective, consider the infant, who, when introduced to something unknown, looks toward an adult in its presence and makes the adult's attitude its own (Hobson 2012:171). Researchers studying this phenomenon have concluded that “the fact that objects at the focus of the [other's] attitudes change in meaning for the infant [depending on what the other's attitudes are, shows how the infant’s attitudes fluctuate in tandem] with the attitudes of the other...such that the world comes to have meaning according to oneself as identified with the other, and therefore...a new meaning for oneself” (Hobson 2012:171). In this way, one’s own representations of objects, events and concepts are formed through interactions with the representations of others. The interactivity upon which our basic abilities to represent the world and our selves as we do depends is what later allows for more sophisticated styles of collaboration. We tend to think that what we feel and know comes about through processes that are largely independent of what others feel and know, but the implications of these findings are that our attitudes, concepts, judgments – indeed, our selves - are not independent constructs but are rather thoroughly inter-relational constructs.

Kindred themes of intersubjectivity were philosophically probed in Martin Buber’s 1923 *Ich und Du (I and Thou)*, in which Buber proffers that “[t]hrough the Thou a man becomes I. That which confronts him comes and disappears, relational events condense, then are scattered, and in the change consciousness of the unchanging partner, of the I, grows clear, and each time stronger...The I is real in virtue of its sharing in reality. The fuller its sharing the more real it becomes." (Buber 1923: 28, 63)"
This ecological approach to appraising selfhood recapitulates an intriguing conceptual turn in the history and philosophy of modern immunology, wherein the paradigm of immune identity as predicated on a strict self/non-self dichotomy has been gradually upended by a paradigm of immune identity that focuses on open and dynamic ecological integration as an organizing principle. Describing this shift in theoretical orientation, Tauber describes that “this a fully ecological perspective (supported by systems biology more generally) [that] alter[s] the basic postulates of immune theory based on an insular self. Instead of a theory grounded on self/non-self distinctions, models of the immune system would be built on an ‘open’ architecture to fully represent the dynamic and dialectical relationship characterizing an organism engaged in its environment… conceptually, the arguments for and against these competing views provide the arena for the philosopher’s own discourse, one that has repercussions for philosophy of biology at large” (Tauber 2008:233-241). The ‘ecological orientation’ elucidated by Tauber bears relevance beyond conceptualization of immune identity to the interdependent architecture of selfhood writ large.

Detailing the relevance of ecological frameworks in immune identity, Tauber (2008) importantly points out that “[w]hile immune cells distribute themselves throughout the body, they are particularly conspicuous at the interfaces between host tissues and the environment: within the skin and underlying mucosal surfaces (e.g., the respiratory tract and gastrointestinal). These are the sites where the body first encounters chemicals and micro-organisms, and thereby senses toxins and destroys pathogens. Such interfaces are obviously open and dynamic, and they possess a complexity distinct to themselves… the organism adjusts its own identity as it responds along a continuum of behaviors to adapt to the challenges it faces, and, indeed, ‘identity’ is determined by particular context. Responses are consequently based not on intrinsic foreignness, but rather on how the immune system sees an ‘alien’ or ‘domestic’ antigen in the larger context of the body’s economy… al. Because such border areas contain species from each habitat, unique forms of competition may occur, giving rise to unique dynamic relationships…. new opportunities arise in such an environment [and] ecotones may be seen as engines of biological innovation. Diversity and dynamism are greatest at the margins between habitats, and it is at such interfaces that whole new biological forms probably originated. Far from being places of pure strife, some ecotones are characterized as much by cooperation and synergism as by cutthroat competition… [an] ecological orientation brings issues of communication and information theory directly onto notions of immune regulation, where different tiers of bidirectional cognition between pathogens and immune cells set the balance of responses and adaptation. Indeed, ‘immune cognition’— replete with metaphorical ‘memory’, ‘perception’, and ‘recognition’—has already provided a new scientific lexicon for a variety of converging conceptual orientations…” (2008:233). See also Tauber, Alfred I., and Scott H. Podolsky. The generation of diversity: Clonal selection theory and the rise of molecular immunology. Harvard University Press, 2000.
This interdependent dimension of selfhood serves as a vital complement to the narrative dimension examined above. Insofar as life stories that persons carry invariably feature others in the social fabric (much as immune identity hinges on the landscape of antigens interfaced with), and conversely, as roles and relationships that exist in a social schema routinely gain full-form only through life-stories conferring meaning and emotional valence to an otherwise inconsequential array of bare facts, the narrative dimension of selfhood is intimately aligned with an ecological orientation.

An interdependent concept of the self carries novel implications for how we think about pathologies of selfhood by shifting the relevant object to include the collective. As such, putative features that may be apparent given a strictly independent self-construal may not be apparent given an interdependent self-construal. Moreover, misconstruing self-concepts at play in human experience may give rise to defective approaches to the diagnosis and management of relevant disorders of selfhood.

**Cognitive Unity**

A remaining key feature to consider in appraising the architecture of selfhood is that of cognitive unity. ‘Cognitive unity’ here refers to a general logical coherence of the attitudes, intentions, motivations, beliefs and volitions that populate one’s mental pace and which enables individuals to effectively deliberate, plan and act in their lives in a manner that is, at least by the agent’s own lights, coherent and reasonable.

This capacity encapsulates more than mere diachronic or synchronic synthesis (which as detailed above, underpins the phenomenological unity of experience) or that of narrative. It refers not to an experiential unity but rather refers to a unity characterized by logical coherence. The role of this capacity is perhaps best illustrated through observation of cases in which it breaks down, such as in dissociative identity disorder wherein an individual might experience events as temporally unified but nonetheless may exhibit a profound logical discrepancy among different beliefs, behaviors and goals (Cf., Itzkowitz et al. 2015). By analogy, a first time player of a complicated game might be able to experience a round of a game as a temporally unified whole, but may lack full understanding or logical execution the rules governing the game, making moves in successive turns that are, with respect to the intended or anticipated outcome, effectively contradictory. The capacity to ensure that items populating one’s mental space are
generally internally consistent (viz., not logically contradictory) allows for a distinctive unity of experience, cognition and action that in part underpins the capacity to live and communicate effectively.

This capacity of cognitive unity is, however, enigmatic, owing in part to a range of philosophical puzzles that emerge when aiming to square this capacity of selfhood with a variety of ordinary phenomena that appear to defy it. Notably, some have argued that making sense of cases of putative irrationality, self-deception and *akrasia* (‘weakness of will’) requires positing a model of the mind that is no longer unified, but is constitutively divided. Sigmund Freud’s extensive observations of clinical neuroses, for instance, famously led him to propose an “architectonic principle of the mental apparatus [that] lies in a stratification, a building up of superimposed agencies” (Freud PE 192). Freud advocated for a “subdivision of the unconscious [as] part of an attempt to picture the apparatus of the mind as being built up of a number of agencies or systems whose relations to one another are expressed in spatial terms…without implying any connection with the actual anatomy of the brain” (Freud PE XX 32). Later theorists, including Donald Davidson and David Pears, adopted alternative views subdividing the mind to meet the conceptual challenges of irrationality, or of cases where individuals appear to act against their own better judgment.

Although mental partitioning may present as an appealing answer to the problems of irrationality, it raises its own array of perplexing issues concerning agency, identity, control, responsibility and the self. Exactly what is it that performs and orchestrates the partitioning? When are these partitions formed, and what is the substrate that is partitioned? If numerous partitions are independently capable of prompting action or belief, is the person to be regarded as a single agent or as a bundle of multiple agents? If as a single agent, what serves to unify the multiple partitions that the agent contains? If as multiple agents, what implications does this have for holding whole persons, insofar as they are collections of agents, accountable for actions brought about by a single recalcitrant partition? Invoking a unifying mental principle or overlapping topology to meet these challenges appears to leave us squarely where we started; viz., having to explain how, in spite of this putative unity, agents can wittingly believe or act inconsistently or against their own better judgement.

Jean-Paul Sartre echoes this criticism in responding to Freud’s early theory of a mental divide between the conscious and unconscious. The problem, Sartre remarks, “still remains of
accounting for the unity of the total phenomenon (repression of the drive which disguises itself and ‘passes’ in symbolic form), to establish comprehensible connections among its different phases. How can the repressed drive ‘disguise itself’ if it does not include (1) the consciousness of being repressed, (2) the consciousness of being pushed back because it is what it is, [and] (3) a project of disguise? No mechanistic theory…can explain these modifications” (Mauvaise foi, 210). “By rejecting the conscious unity of the psyche,” Sartre proceeds, “Freud is obliged to imply…a magic unity linking distant phenomena across obstacles… [yet, this] explanation by magic does not avoid the coexistence – on the level of the unconscious, on that of the censor [which unifies the parts], and on that of consciousness – of two contradictory, complimentary structures which reciprocally imply and destroy each other. We [thus] find that the problem that we had attempted to resolve is still untouched” (Mauvaise foi 210-11). In explaining how partitions unite within one person to dynamically interact, one is confronted with the same problem motivating theoretical efforts to partition the mind in the first place. Pears cites this criticism in Motivated Irrationality, but does not explain how his formulation of mental partitioning resolves the problem. In Freud’s Anatomies of the self, Thalberg raises a related host of “identification problems” with Freud’s treatment of the mind (254), and has us “[r]ecall Freud’s treatment of our instincts or drives. Freud says ‘these processes strive toward gaining pleasure; psychical activity draws back from any event which might arouse unpleasure’ (1911b, XII, 219). But whose enjoyment do they ‘strive’ for? Why mine? Surely it is unintelligible to suppose they enjoy escaping from my homeostatic mental apparatus…. [w]hich ‘self’ does my ego have the duty of preserving? How is its continued existence related to mine?...What exactly do we mean by ‘its own advantage’? [The] notion of this prominent actor within us…seems quite elusive” (254). The problems articulated by Sartre and Thalberg are perhaps most pronounced when considered through the lens of Freud’s theory of the mind, but it is not clear whether more contemporary theories of partitioning successfully avoid them.

A few key points gesturing toward a solution to this problem of how to reconcile the conceptual need to posit some form of cognitive unity with the range of regular instances of human action and cognition which seem to defy it should be noted in defense of cognitive unity as a durable dimension of selfhood.

Theories of mental partitioning in accounting for irrationality and akrasia stem from a philosophical impulse to preserve a certain orthodoxy about action according to which every
action, by virtue of being an action, is perforce caused by a primary reason which stands to explain and to motivate its behavioral effect (see, for instance, Davidson 1980 in *Actions, Reasons and Causes*). On this view, an agent’s behavior at any point in time only constitutes action if and only if it is caused intentionally by a certain kind of reason-oriented mental state. This generates difficulties when trying to account for irrational action, requiring us to either jettison its conceptual possibility (a view dating back to Socrates, who famously professed, “no one voluntarily pursues that which he judges to be wrong” (Protagoras 358b-d)) or to preserve its possibility by positing mental partitions containing competing sets of potentially motivating reasons.

Contrary to this standard story according to which all actions are, perforce, motivated by primary reasons, it may well be the case that actions are only contingently motivated by primary reasons. In cases where they are, such actions could be deemed rational, and in cases where they are not, such actions could be deemed not rational (arational or irrational). For perhaps what it means to be irrational is to act without any primary reason at all, where one could have acted with a primary reason.10

---

9 The translation here is in line with Norman Gulley’s in *The Interpretation of 'No One Does Wrong Willingly' in Plato's Dialogues* Phronesis. Vol. 10, No. 1 (1965), pp. 82-96

10 On this counterfactual account of irrational action, behaviors are bisected into two main classes: (1) behaviors prompted by primary reasons (2) behaviors not prompted by primary reasons. On this approach, all behaviors of the first class would be rational actions. That is, every member of the set BMP (every behavior prompted by a primary reason) would meet the conditions necessary to qualify as a rational action. In turn, behaviors of the second class could fall into two subsets: (iia) behaviors not prompted by primary reasons where the agent could have acted with a primary reason (iib) behaviors not prompted by primary reasons where it is not the case that the agent could have acted with a primary reason. While actions of the first sort (iib) would be non-rational behaviors (such as muscular twitches and compulsion), behaviors of the second sort (iia) would be classified as irrational actions. Hence an agent’s action would be deemed irrational (IA) if and only if an agent acted without a primary reason in a case where the agent could have acted with a primary reason. Behaviors of kinds (iia) and (iib) are alike in that they lack causal connections to primary reasons, but differ in respect to the state of the agent when each type of behavior respectively occurs. Agents performing actions of type (iib) neither are moved to act by a primary reason nor have potential primary reasons at their disposal pursuant to the occurrent behavior or its negation. If these pages fall to the floor, sending a loud thump through the room causing the reader to flinch, the reader’s flinching would be considered a behavior of type (iib). With neither a primary reason causing it nor a reason at the reader’s disposal relevant to flinching or ~flinching in this scenario, it is not an action at all. Agents performing actions of type (iia) are markedly distinct. Although they act without primary reasons to rationalize what they do, they do have reasons at their disposal relevant to what they fail to do. By virtue of having such latent primary reasons and failing to use them, such agents would on this account be considered irrational. This sort of counterfactual theory of irrational action fits well with many of our intuitions about what irrational actions are, and seems more parsimonious than the theory of action...
On this view, when an agent acts irrationally, there is no primary reason that needs to be sequestered from the agent’s better all-things-considered reason. Whereas the standard account designates primary reasons as necessary conditions for all actions, this account delegates primary reasons to the more circumscribed conceptual role as sufficient conditions for rational actions. Of course, this account does not entail that mental partitioning is conceptually impossible; instead, it is explains why mental partitioning is not strictly necessary to explain irrationality and that there are alternative ways of viewing action and agency that do no require a constitutively partitioned mind. Moreover, this account does not rule out the possibility that the irrational mind only sometimes is partitioned, or that sometimes mental partitioning does account for irrational behaviors. Rather, it has sought only to explain why the concepts of agency and action themselves do not conceptually oblige us to partition the self in accounting for the colorful array of conflicts that populate the economy of action and thought.

III. Pathologies of Selfhood

The foregoing analysis reveals that our concept of selfhood encapsulates a wide repertory of symbiotic capacities. To review, seven key capacities that might be said to be constitutive of an intact scaffolding of selfhood have been elucidated:

1. Diachronic unity
2. Synchronic unity
3. Self-other representation
4. Narrative (past, present and future)
5. Consciousness (in particular, awareness)
6. Ecological embeddedness
7. Cognitive unity

supported by Mele, Heil, Davidson and others. The word ‘irrational’ derives from and is defined as “not governed by reason” (Webster). It does not mean “not governed by best reason” but rather refers to action devoid of reason altogether. This account captures this common meaning well while obviating the need to partition the mind. Indeed, if primary reasons are entirely absent in cases of irrational action, then there is no discernible logical need for partitioning.
Though not meant to be exhaustive of the entire conceptual terrain, these elements might be regarded as key capacities that coalesce to form an intact scaffolding of selfhood.

This approach represents a marked departure from prevailing approaches to selfhood; rather than reducing the self to a monotonic concept with a single common denominator, the self on this account consists in a network of crisscrossing symbiotic capacities that share some family resemblance and displays a characteristically open texture. “[A]s in spinning a thread,” to borrow the apt description of family-resemblance concepts Wittgenstein invokes in describing our concept of numbers, “we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres” (PI §67; Cf., Waismann 1951).

Substantial alterations of any of these elements could foreseeably result in a fracturing, disintegration or transformation of the emergent self, resulting in an array of conditions, from the ordinary to the pathological. The phenomena of thought-insertion delusions in schizophrenia and mirror-touch synesthesia as explored in preceding chapters provided examples of conditions emerging with disturbances of self-other representation. Also briefly considered above was dissociative identity disorder, manifested by a loss of cognitive unity.

Indeed, variable disturbances of each of the capacities enumerated above may result in distinctive clinical presentations, furnishing a rich empirical basis for studying the phenomenology and neurobiology of each respective dimension of the self. Accordingly, conceptualizing and studying such conditions qua disturbances of selfhood may help to reveal novel features of their phenomenology, and accordingly illuminate novel areas for future research and guide approaches to therapy.

With respect to the diachronic dimension of selfhood, striking cases to consider are those of chronotaraxis (‘time amnesia’) and chronoblindness, pronounced confusion with respect to time, or inability to perceive the passage of time. In such cases, patients may experience acute or prolonged temporal disorientation. These conditions have been described as being occasioned by thalamic lesions (Cf., Spiegel et al. 1956, Kumral et al. 2007, Lee et al. 2010) or right parietal pathology (Cf., Rocha et al 2015). In their supplemental data, Lee et al. report the fascinating case of this sort:

A 38 year-old gentleman was “admitted for a sudden somnolence while playing golf [in the fall of 1996]. When admitted to the hospital, he was in a state of stupor without other focal neurological abnormalities. After MRI and a full work-up, he was diagnosed with the
bilateral medial thalamic infarction in the area of the paramedian thalamic arteries arising from a common trunk… His consciousness returned after a week, although he presented a severe impairment of time orientation. This impairment has continued for more than 13 years… he was unaware of the day, month, year, and exact time at the interview. He was also unaware of the duration of his stay in the clinic room, the time taken from his house to the hospital, and even the ages of his children. When asked the age of his 21-year-old son, he answered that he was 7 years old, his age in 1996. He was able to recognize the Vancouver Winter Olympic Games and the name of the national medalists, but failed to recognize the month and year (February, 2010) when the Olympic games were held. He also remembered the names of the coaches and players of Korea’s national football team and even knew the scores of the FIFA World Cup held in Korea and Japan in 2002, but he did not know the year that the games were held. He recognized all the Korean presidents when given the names (10 presidents since 1948); however, he did not know the order of the presidents who were elected after October 1996, although he correctly identified the order of the presidents before 1996. Nevertheless, he was able to accurately identify periods of 10 or 20 seconds and had good musical rhythm, suggesting that his internal clock was normal. He had also normal biologic rhythms, including a normal sleep cycle. In other neuropsychologic examinations, he had a normal performance in tests of episodic, semantic, and implicit memories, as well as in tests of attention, praxis, frontal executive functions, frontal inhibitory function, visuospatial function, calculation, and language. After the symptoms began, he failed to return to his former job … Although the patient remembered all events after his stroke, they were registered without time information. Thus, although the patient seemed to have normal episodic memory, he had severe difficulty in remembering the time information associated with events and in retrieving them in the appropriate sequence” (Rocha et al. 2010, supplemental document A1, page 2-3).

Cases such as this reveal the importance of diachronic unity in maintaining higher order functioning, while at the same time revealing that its isolated loss does not entail the wholesale loss of selfhood (a conclusion which, it might be argued, may be entailed by a minimalist Kantian conceptualization of the self). It moreover provides insight into the possible neural underpinnings of this particular dimension of selfhood as diachronic unity, narrowing potential localization to the thalami and parietal lobe. Studying instances of simultagnosia (inability to perceive more than a single object at a time), akinetopsia (inability to perceive motion (Cf., Cooper & O’Sullivan 2015)), and related attentional disturbances to clarify the neural underpinnings of synchronic unity in a similar fashion.

Turning our attention to pathologies of the narrative dimension of selfhood, it is apt to consider how loss of autobiographical memory in Alzheimer’s disease may characteristically give way to a distinctive loss of narrative, and yet other dimensions of selfhood may remain
intact. As such, conceptualizing Alzheimer’s disease as an en-bloc “loss of self” (Cf., Cohen & Eisdorfer 2001) is misconceived. Delicate care should be taken among clinicians and researchers in deciding how to conceptualize and explain alterations of self in such common clinical conditions, the communication of which could impact not only how patients and families come to make sense of their condition, but may also bear significantly on debates in public policy surrounding patient rights and legal norms, as may be gleaned from the recent debate in the Netherlands regarding euthanasia for patients with dementia (Cf., Hertog et al. 2007).

A nuanced approach that carefully delineates capacities altered or lost without withdrawing attention from the dimensions of self that are preserved (and may even continue to be enriched) is imperative. Indeed, alterations of some dimensions of self may be accompanied or succeeded by the emergence of other dimensions of self; Miller et al. (1998) describe a case series of patients who became visual artists in the early stages of frontotemporal dementia, and explain how such cases provide unique windows into studying creative aspects of selfhood. Cases such as these challenge theoretical orientations conceptualizing selfhood as a monotonic or static.

Among the cases Miller et al. describe is a 68-year-old man who “was seen for a dementing illness of a 12-year duration. Previously a successful businessman without interest in art, at age 56 he began to describe “open” and “closed” periods. When “closed,” he was dysphoric, and experienced lights and sounds as exquisitely intense. When “open,” lights and sounds produced a pleasant feeling that allowed him to think creatively. He painted images experienced during “open” and “closed” periods. At 58 years he became anomic and disinhibited. Language and memory deteriorated, but he showed heightened visual and auditory awareness. Odd compulsions developed, and despite his considerable wealth, he cajoled his caregivers to walk with him to look for coins. At 56 years he began painting. During the next decade he created paintings with increasing precision and detail. The first featured brightly colored ellipses. Soon his work became realistic, and he drew animals. Later works were crafted with care, and he took hours to complete single lines. Between 63 and 66 years his paintings won several art show awards. … He displayed heightened interest in his environment, commenting extensively on color and sound…” (Miller et al., 1998:980). The emergence of creative

11 I am grateful to Dr. Bruce Miller and Dr. Hannah Kirsch for bringing these cases to my attention during my recent visit to UCSF.
capacities has more recently been described in cases of other neurologic diseases, including verbal creativity in transient epileptic amnesia (Woollacott et al. 2015), musical creativity in some forms of cerebrovascular disease (Zagvazdin 2015), poetic talent in Parkinson’s disease (Inzelberg 2013), and artistic ability in traumatic brain injury (Midorikawa & Kawamura 2015).

Other conditions to consider in evaluating the range of cases that might inform and be informed by our schematization of selfhood include transient or persistent amnestic syndromes arising in the context of epilepsy, autoimmune neurologic conditions, focal lesions, and in cases of psychogenic or trauma-induced amnesia and fugue (Harriot 2015; Campora 2016; Frisch 2013). In such cases, loss of episodic memories might entail disturbances of selfhood insofar as the capacities for diachronic unity and narrative are implicated, in addition to possible disturbances in self-other representation. Felt disturbances of selfhood occasioned in the context of amnesia are reflected in the phenomenology of some reported cases (Cf. Nochi 1998), but most published accounts lack sufficient phenomenological detail to parse out which dimensions of selfhood patients and clinicians might be referring to when describing a loss of identity or a dwindling away of self.

Strikingly, apparent disturbances of selfhood in patients with Anton–Babinski syndrome (visual anosognosia wherein cortically blind patients do not recognize their visual deficit), and other forms of anosognosia (unawareness of deficits) and somatoparaphrenia (a sense of alienation from parts of one’s own body) arising in the context of cerebrovascular accidents, traumatic brain injuries, and in neurodegenerative disease may manifest not only with disturbances of self-other representation but also with disturbances in cognitive unity; interestingly, the therapeutic technique of presenting the self to the anosognosic or somatoparaphrenic patient from a third-person perspective (e.g., through mirrors or videos) has been found to be effective at reinstating motor and proprioceptive awareness in some patients (Fotopoulou et al. 2009, 2011). On the basis of these findings, Fotopoulou et al. suggest that such conditions “can be regarded as a neurogenic dissociation between the ‘subjectively felt’ and ‘objectively seen’ body. This recalls the developmental finding that young infants cannot link their ‘felt body’ with the view of themselves in a mirror,” implying that there may be multiple neural networks at play responsible for simultaneous representations of the self from different perspectives, in addition to integrative network(s) unifying these discrete representations under
normal circumstances (2011:3946), generating a large range of potential pathologies of selfhood that might ensue owing to breakdown at any point within these complex networks.

Beyond the range of cases already considered, cases of commissurotomy (and the accompanying ‘split-brain syndrome’ (Cf. Schechter 2015)), disorders of consciousness, schizophrenia beyond thought-insertion, and other neuropsychiatric syndromes provide rich and relatively underexplored opportunities to further expand and crystalize our understanding of the profound phenomenology and neurobiology of selfhood and its multiform pathologies.
Future Directions

The analysis offered here illuminates numerous avenues for future research, some of which have already been gestured at above. These include fruitful avenues for both empirical and conceptual work.

On an empirical plane, further research exploring the neural substrates corresponding to different components of the scaffolding of selfhood articulated herein, as well as how these varied neuronal systems/networks coalesce to produce a phenomenologically integrated and unified self is much needed. Further clarity with respect to these issues can shed further light on how particular brain lesions may differentially affect elaborations of selfhood by impacting different dimensions of selfhood. Such research could foreseeably include functional neuroimaging studies, non-invasive brain modulation (e.g., TMS; tDCS), and targeted neuropsychological testing of individuals with apparent disruptions of self stratified according to the domains of diachronic unity, synchronic unity, self-other representation, narrative, awareness and interdependence to aid in elucidating the enigmatic neuroanatomy and neurophysiology of selfhood. Moreover, studies examining the emergence and development of each of these dimensions during key windows of human development can serve to clarify the ontogeny and organization of selfhood over the lifespan. In the pursuit of such research, it is important to avoid a misleading philosophical impulse toward mereological reduction with respect to the self, and to aim for utmost conceptual clarity in the construction and operationalization of sound methodologies that effectively articulate which feature(s) of selfhood or identity are being measured and how.

On clinical and conceptual planes, there is foreseeable utility in future work geared toward developing a quantifiable, multidimensional index of selfhood to assess patients’ robustness of self through measurements of the various capacities enumerated herein; such a clinical tool could help to evaluate and longitudinally track patients’ development through time.

12 P.M.S. Hacker and others have famously argued for what is dubbed the mereological principle - that psychological properties can only be justifiably ascribed to persons as whole beings and not to constituent parts of persons. This principle is wielded to contend that many contemporary neuroscientists and their respective methodologies are guilty of a mereological fallacy in ascribing psychological properties to parts of an individual instead of to the individual as a whole. This is not to say that components of the individual cannot cause or underpin psychological predicates applying to a person as a whole, but rather that it is the whole person and not a part of the person that instantiates and experiences the psychological predicate of interest (see Hacker & Smit 2014 for more on the mereological fallacy).
in settings of disturbances of selfhood, according to some of the measures that might matter most to patients and families. Such a tool might be of further utility in future clinical studies evaluating the natural history of neuropsychiatric diseases and in studying novel methods for their diagnosis and prognostication.

Additionally, future research appraising the adequacy of current neuropsychiatric nosologies in capturing the breadth and depth of pathologies of selfhood may yield actionable insights. Wittgenstein’s family-resemblance approach may serve as an ideal conceptual framework for optimizing current taxonomies of neuropsychiatric disease.

Finally, the relative paucity of existent literature detailing the experiences and attitudes of patients and families around disorders of self can be addressed through the compilation of an organized compendium of clinical cases detailing the diversity of pathologies affecting selfhood and elucidating their respective phenomenologies through targeted interviews of patients and families. Inter alia, such a project could serve the further aim of elucidating optimal communication styles and interview techniques that clinicians may adopt during patient encounters and when discussing the conceptually challenging issues of selfhood and identity with patients and families. These are projects that I plan to undertake moving forward.

The cross-disciplinary nature of each of these avenues of future work underscores the profound importance of deep and sustained collaborations among philosophers, clinicians and researchers to optimize and advance approaches to the diagnosis, evaluation and treatment of pathologies of the mind and self.
Summary and Conclusions

This Thesis has examined the crucial yet underexplored convergence of philosophy and neuropsychiatry with respect to selfhood and first-personal authority, particularly through the lens of thought-insertion delusions and mirror synesthesia, and illuminated a novel account of selfhood emerging at this intersection.

Chapter 1 demonstrated how the psychopathology of thought-insertion delusions can inform our philosophy of mind, and how philosophical insights may in turn help describe the nature of thought-insertion (TI). It elucidated how the phenomenon of thought-insertion may be reconciled with the immunity principle (IP), a canonical principle in philosophy of mind according to which first-person, present-tense reports about psychological states are immune to errors of identification, and revealed why it is crucial that any formulation of the immunity principle be limited to present-tense self-referential judgments ascribing some predicate to the essential indexical ‘I’ based on a privileged first-personal perspective. It concluded by examining why thought-insertion remains puzzling even if it does not challenge any obvious philosophical principle.

Chapter 2 explained and critically evaluated arguments raised in response to my accounts of IP and TI, and examined possibilities for non-delusional first-person-like awareness of mental states that are not one’s own through exploring the condition of mirror synesthesia. It concluded that instances of mirror synesthesia blur the boundaries between identification and misidentification, and challenge philosophical orthodoxies about the self in a manner intriguingly similar to TI.

Chapter 3 situated foregoing insights about self-other representation into a broader account of selfhood, explaining that the capacity for self-other representation is not wholly constitutive of selfhood. It delineated and defined six other capacities typically present an intact scaffolding of selfhood, situated this project in the history of ideas, and examined how substantial alterations of any of these elements of selfhood could result in a fracturing, disintegration or transformation of the emergent self. Future appraisal of how particular lesions may differentially affect elaborations of selfhood by impacting neural substrates and networks corresponding to different dimensions of this complex scaffolding was explained.
It concluded by illuminating avenues for future research, and by urging sustained collaboration between philosophers and physicians to optimize and advance approaches to the diagnosis, evaluation and treatment of pathologies of the mind and self.

It is my hope that the ideas developed herein may pave the way for more accurate criteria for describing, diagnosing and managing pathologies of selfhood, and may offer insights to theoreticians, clinicians and patients alike in search of clarity about the self and its pivotal place in health and disease.
References


Brent, Benjamin K., et al. 2014. Self-disturbances as a possible premorbid indicator of schizophrenia


Gulley, N. 1965. The Interpretation of 'No One Does Wrong Willingly' in Plato's Dialogues. Phronesis. 10.1:82-96


Hacker, P. M. S. 2015. An Intellectual Entertainment: The Nature of the Mind. Royal Institute of
Philosophy Supplement 76: 199-223.
Hofmann, Björn. 2001. Complexity of the concept of disease as shown through rival theoretical frameworks. Theoretical medicine and bioethics 22.3:211-236.
Press


Appendices

Appendix 1a: Dr. Clara Humpston’s Forthcoming Commentary to which the analysis in Chapter 2 in part responds
In his paper, Young has eloquently put forward a novel account of how and why the phenomenon of thought insertion (TI) seen in patients with schizophrenia does not contradict the immunity principle (i.e. first-person access to one’s own thoughts and psychological states is ‘immunity to error through misidentification’). He argues that in TI, the problem lies not in misidentification but in misprediction: the individual with TI does not ascribe the right predicate to the wrong subject but has misdetected the predicate in the first place. The author points out that an inconsistently formulated immunity principle could risk confusing the two types of errors. The author defines the immunity principle as:

\textit{Any agent who thinks a thought}^* \textit{(the asterisk indicates thoughts obtained by privileged access)} \textit{of the form ‘I am F’ cannot be right on the basis of the perspective grounding this thought that something is F but wrong about whether ‘I am F’}. 

Given this definition, it would indeed be the case that if one claims they have thoughts that are not theirs, it is the predicate and not the initiator of the thoughts that is susceptible to error. Note that the author stresses ‘any formulation of the immunity principle be limited to present-tense self-referential \textit{judgements} [my emphasis]’ ascribing a predicate to the essential indexical ‘I’, and any over-simplification would only create the illusion that TI breaches the immunity principle.

However, perhaps another possibility for this confusion is the inconsistency in defining TI itself. The author of the current paper makes a specific point that in TI, the patients do not
only claim the thoughts are not generated by themselves but actually belong to another agent. This, of course, would seem implausible (whether considered from the immunity principle or not) if not utterly false even to a lay person; and it is this apparent falseness that defines TI, at least in this instance, as a delusion. Indeed, throughout his paper Young has consistently referred to TI as ‘delusions of thought insertion’ and implies that the delusion is primary in nature (‘patients seem to require no external evidence, investigation or complex inference in order to arrive at them [the delusions]’). This conforms to the psychiatric definition of TI as a ‘false belief that the subject receives inserted, alien thoughts’ (Mullins and Spence 2003).

Arguments about whether all delusions are beliefs aside, this definition could be divided into two components: the false belief and the inserted alien thoughts. Just by this medical definition alone, it is clear that the belief part has been dissociated from the actual inserted thoughts. So, if TI is indeed a delusion, to which of these two components does the delusion apply? It would seem obvious that the false belief about the inserted thoughts is the delusion. Some may argue that how could thoughts be inserted from the external (and also belong to someone else) in the first place? ‘Privileged access’ allows one to have direct, taken-for-granted knowledge (a kind of knowledge which does not require external evidence) of one’s own thoughts which are obtained through introspection. Yet, thoughts derived from both introspection and the mere act of thinking are separated from the individual’s subjectivity in TI. This separation (some call it the ‘separability thesis’, see Gibbs 2000), I argue, is the experience of having seemingly external thoughts. The individual who has such an experience does not deny the fact that they are the one who has had external thoughts inserted into their mind; the individual simply denies the thoughts themselves are their own. Indeed, this would fit nicely with the author’s account about mispredication. However, does this always constitute a delusion?

I have recently proposed (Humpston and Broome 2015) that the experience of having external thoughts – whose externality is only accessible through the first-person perspective (which is again a paradox) – alone is insufficient for a delusional elaboration to form, and it is the act of ascribing another agent to these thoughts that drives the individual to a delusional endpoint of the phenomenon. I do not necessarily agree with the term ‘delusions of thought insertion’ but rather, I think it is more likely ‘delusions in thought insertion’. Certainly the difference of one proposition may sound like a deliberately difficult game of lexicon, but I would like to argue that it is an important distinction which highlights the duplex nature that
‘completes’ the phenomenology of TI. In this sense, Young’s account about mispredication would only apply to the basic, generative *experience* of TI and the delusions in TI would in fact be susceptible to errors through misidentification because the subject has actually ascribed the wrong instantiator to some thoughts that are in their own mental space. For example, if someone has an inserted thought of ‘KG’ (to use a well-known case), the thought itself does not contain ‘I must KG’ but the indexical ‘I’ comes as a given prerequisite of thinking. Therefore, when the patient ascribes this thought to ‘Chris’ (again to use the author’s example) they have wrongly identified the agent of such ascription. Without this process of ascription or judgement, the delusion would not form and neither would the immunity principle apply.

It is true that this distinction between thoughts and thinking may appear to be trapped in an infinite regress (‘believing a thought is a thought is a thought’, and so on). Nevertheless, I think it could be defended by a more detailed and precise analysis of the ‘agent’ in TI. The impression from Young’s paper is that he deals exclusively about the ‘judgement’ of agency (hence my emphasis at the beginning of this commentary) which to me is a second-order process involving the attribution of a causal agent to a particular set of thoughts. The *sense* of agency, on the other hand, would be much more pre-reflective and basic. Young writes that ‘patients do seem to *arrive at* [my emphasis] their delusional thoughts through a distinctive, first-person way of knowing’. What he does not detail is how the patients could obtain this knowledge through, say, privileged access yet engage in an attributional process of the (seemingly) same agent to an external body simultaneously. Of course, as mentioned previously, delusions in TI are primary in nature in Young’s account – one which overshadows the entirety of the experience – but to me the delusional elaboration is secondary for reasons discussed above. While patients may not require any ‘complex inference’ (but if all delusions are doxastic surely they would require some sort of reasoning to gain belief status?) to find an external agent, it seems that the ascription of thoughts to this particular agent (e.g. ‘Chris’) is a parallel and dissociable process from the kind of first-person access that is endowed with immediacy and transparency. The agent of awareness has been detached from the agent of authorship and ownership when the patient ‘hands over’ their thoughts to someone else. In TI, the latter is severely impaired or absent whereas the former may still be relatively intact (certainly the patient is still consciously aware of *any* thought that is in their mind!). This may answer Young’s question of ‘why the patient is delusional in the first
place’ – because the patient is not delusional at first when they experience the ‘inserted’ thoughts but *becomes* so when they arrive at the conclusion that they are thought by an external thinker.

I understand some might question how it is at all possible to have thoughts that are generated by someone else in one’s own mind; isn’t this somewhat too concrete and bordering on solipsism? Yet in my opinion such an orientation captures the exact experience of the patient’s way of thinking and indeed, even ‘normal’ ways of thinking can sometimes be viewed as solipsistic as they reflect the very nature of subjective reality created by the experiential field of each and every one of us, delusional or not.

Agreeably, just as Young himself writes, there is something deeply mysterious about the phenomenon of TI no matter how much one attempts to explain it with whatever kind of theory. TI itself may not challenge the immunity principle as such, but it still threatens the continuity and integrity of the subjective existence of thinking and the thoughts that ensue. Due to the ‘lack of robustness of one’s own self-boundaries’, the psychotic individual is trapped in a perpetual cul-de-sac, a paradox that erodes their very being. At the same time, however, they still use the first person pronoun ‘I’ to describe their bizarre and often extremely distressing experiences. This ‘I’ may not be the same as the essential indexical ‘I’ in this context because there is simply no other replacement for ‘I’ of any kind. Patients use ‘I’ as a mere descriptor which lacks the everyday mineness and associated immediacy others take for granted.

I have also argued that self-boundaries are not all-or-none concepts, but in fact lie on a continuum between internality and externality (Humpston and Broome 2015) and are permeable for subjective thoughts and experiences to ‘diffuse’ across. Of course, one could very well argue that all externality in this case is in fact internality because how could thoughts occur without a thinker? The counterargument is that while thoughts do not occur without a thinker, they can still lose the status of having an original thinker after they have been thought, especially when the original thinker themselves (mis)attribute them to another agent. There are many potential reasons why the original thinker may (consciously or unconsciously) decide to do so, or it could even be an automatic process to expel thoughts that in one way or another threaten the stability of the individual’s subjectivity. However, this creates yet another paradox, because for such diffusion to occur the subjectivity must have been already fragmented so there would not be a stability to protect. Perhaps the sheer pain and suffering caused by psychosis (with TI as a first-
rank symptom) are the exact manifestation of these unresolvable paradoxes originating within one another.

In sum, Young’s compatibilist account of TI and the immunity principle is compelling, but might benefit from a more detailed re-analysis of the distinct types of thinking and agency of thought as well as a clearer definition of the prerequisites under which the argument stands. Here I have outlined some of the implications and theoretical considerations of TI which will undoubtedly remain a mystery. I agree with the author to not risk oversimplification in one’s pursuit of philosophical generality – perhaps accounts of TI can never be generalised anyway – but I think it is important to delineate to the best of one’s ability the manifold complexities of the puzzle that is TI.

References


Appendix 1b: Dr. Lauren Ashwell’s forthcoming commentary to which the analysis in Chapter 2 in part responds
Possibilities of Misidentification

Lauren Ashwell
2 Andrews Rd, Lewiston ME 04240 USA
lashwell@bates.edu
+1 617 642 2381 (phone)

We appear to have a special, seemingly direct relationship to our own thoughts that we don’t have to the thoughts of others; I can become aware of my thoughts in a way that I can’t become aware of yours: through introspection. Those who have delusions of thought insertion, however, claim not only to be aware of another’s thoughts, but to have another’s thoughts in their own mind. These thoughts, of course, cannot actually be someone else’s thoughts. However, if we take those who have this delusion at their word, it certainly seems to them that these thoughts belong to another. While introspection is no longer generally thought to be absolutely error-free, this seems quite a strange mistake to make. Surely we can’t misidentify who has the thoughts we are aware of through introspection! Yet the existence of thought insertion delusions shows that this kind of misidentification is not just possible – it actually happens. So thought insertion delusions seem to threaten the principle of Immunity to Error through Misidentification (IEM) of our introspective mental state attributions (Campbell, 1999a). However, in “Pathologies of Thought and First-Person Authority,” Michael Young (2016) argues that thought insertion delusions are not counterexamples to this immunity principle, once we properly formulate it.

One might gloss IEM by saying things like, “we can’t misidentify who has the thoughts that we are introspectively aware of,” just as I did above. And thought insertion delusions, if we interpret these reports straightforwardly, certainly are counterexamples to this. Yet, as Young points out, they are not counterexamples to the following principle:

Immunity Principle (IP): any agent who thinks a thought* (where thought*s are thoughts, based on introspection, attributing a mental state in the present-tense) of the form ‘I am F’ cannot be
right on the basis of the perspective grounding this thought that something is F but wrong about whether ‘I am F’.

If IEM for thought*s is correctly characterized by IP, then Campbell has provided a counterexample, for the principle was only ever meant to apply to first-person ascriptions of mental states (which Young calls “I-thought*s”) – those judgments wherein you ascribe the mental state to yourself through introspection. When suffering thought insertion delusions, patients ascribe the mental state in question to someone else. Moreover, Young argues, if we reformulate descriptions of thought insertion delusions so that they are in the first-person (so the patient’s ascription denies that the mental state in question is theirs), we end up with a different sort of error – an error of mispredication. That is, we simply end up with the patients making a mistake about which mental state they are in (or are not in), rather than an identification error.

While IP can be defended from counterexample in this way, I will argue that IP doesn’t show us anything about introspection or the first-person – which should make us wonder whether it really captures what’s at stake in discussions of IEM. First, Young’s defense of IP turns on features that are not peculiar to introspection or the first-person – nor to any other kind of subject matter that is thought to have IEM. Second, although Young has given us a principle that actual-world thought insertion delusions do not threaten, IP is not yet strong enough to capture the standard claim of IEM for introspective self-attribution. Without further supplementation, both IP and Young’s defense of IP leave open the possibility that such immunity might be merely contingent. Furthermore, thought insertion delusions provide prima facie evidence that IP is merely contingently true.

The general structure of Young’s argument can be used to defend a multitude of versions of immunity to error through misidentification principles – not just ones that involve first-personal introspective judgments. For example:

(BIP) Any agent who thinks a thought based on evidence concerning Bates College of the form ‘Bates College is F’ cannot be right on the basis of that evidence that something is F but wrong about whether Bates College is F.
Of course, someone could think that *Bowdoin College is in Lewiston* on the basis of evidence about Bates – perhaps a visit to the Bates campus in Lewiston, and seeing someone, who is mistaken for a student at that college, wearing a Bowdoin t-shirt. Isn’t this a counterexample to BIP? Not so, for BIP only applies to statements about *Bates*. But what if we reformulate the counterexample so it *is* about Bates: “Bates is not the college in Lewiston”? This would be just a *mispredication*, and so not a counterexample to BIP. This defense of BIP turns on a stipulation that the grounding evidence is evidence about *Bates College*. So long as we have that stipulation, BIP can be defended in exactly the same way Young defends IP.

This is not to say, however, that there are no differences between BIP and IP. First, IP does not talk about evidence, but instead about grounding perspectives. However, I suspect that Young talks of grounding perspectives so IP is silent on whether introspection proceeds by evidential or evidential-like reasoning. Second, IP does not exactly stipulate that the grounds for self-ascription must be one’s own mental life, but instead stipulates a *way* of coming to those grounds. Yet it is only contingent that other subject matters lack such an exclusive “way” to obtain grounds for judgments about them; we can surely imagine having a faculty whose exclusive use is for coming to judgments about Bates College, although such a possibility is definitely a strange one. And in that possibility we could single out that “way” of gaining grounds, and build a principle that parallels IP extremely closely.

It may be thought, though, that there is another important difference between these cases: in my example, surely one needs have extra information that the method being used is exclusive for knowing about Bates College; in the first-personal case, one does not need this extra information, for “the information that makes it sensible for one to think that there is some property being instantiated concomitantly individuates the thing that instantiates the property” (Young 2015, 7). Yet IP, and therefore BIP, only mentions being right and wrong in identification – we need not worry about whether the thinker has information that the method *is* an exclusive way to come to judgments about Bates College, only that she uses it to make a judgment about Bates College. It may be that to *know* that Bates College is in Lewiston simply on the basis of the campus visit one also needs to know the identifying information that *this* college is Bates College, and perhaps one does not need this in the first personal case (I will discuss this further, below), but this difference is irrelevant to Young’s defense of IP, and my analogous defense of BIP.
So far I have argued that Young’s defense of IP is not grounded in anything necessarily peculiar to introspection or the first-person; it rests on the fact that introspection delivers information only about one’s own mental states, and introspection could share this feature with other methods of belief-formation. This is surprising. Although first-personal present-tense mental state attributions are not the only kinds of judgments deemed to have IEM, one would think that the interest in IEM for the first-person is grounded in its relevance to the nature of introspection – or at least in something interesting about indexicals.

One might object that introspection’s exclusivity is more special than any example I can cook up – isn’t introspection necessarily exclusive? It is often said that IEM (of I-thought*s) is a logical principle; indeed, Young refers to it as such several times. Sydney Shoemaker (1970) distinguishes de facto IEM – where the immunity arises from contingent features of the situation – from logical immunity to such error. So IEM for I-thought*s, as Shoemaker intends it, is not based on a merely contingent matter of introspection’s first-personal exclusivity. If it is possible for introspection to provide awareness of another’s mental state to mistakenly attribute that state to oneself, then, in that possibility, IP would be false.

One way to defend the claim that introspection is necessarily exclusively first-personal is to define introspection as involving access to only one’s own thoughts. But while this makes the principle unassailable, it also makes it uninteresting. We don’t say anything substantive about introspection or the first-person if this matter is to be settled by fiat – only about how we choose to use the word “introspection.” Presumably IEM is more noteworthy than this.

Whether introspection could deliver information about others’ mental states isn’t settled without some detail about how introspection operates; if Young intends to defend this principle as necessarily true he should not give the disclaimer that he will “not […] explicate the mechanism(s) by which this way of knowing operates” (Young 2015, 4). Perhaps we don’t need a complete positive account, but to defend logical IEM we certainly need more. Annalisa Coliva (2002) puts forward a similar argument to Young’s, but builds in an assumption about how introspection proceeds: that introspection of a particular mental state requires being in that very token mental state. If we grant that assumption, then I-thought*s would indeed have logical IEM. Shoemaker’s own account of introspection holds that the introspective awareness of the mental state ‘contains’ the mental state introspected, and so rules out the possibility that I might be introspectively aware of a mental state that I don’t have. Similarly, some forms of transparency
accounts of introspection require that one must be in the mental state being introspected (see, for example, Byrne 2005).

Yet these accounts of introspection make our introspective access, at least where we positively attribute a mental state, infallible. We should not require that (positive) introspection is infallible in order to argue for logical IEM. Even this is tempting for introspective awareness of thoughts, it seems far too strong for other mental states, like desires. And if we accept this even just for thoughts, we need to be careful in stating the view if we wish to hold that reports of thought insertion are meaningful, as Young suggests he does.

Another issue with supplementing Young’s argument with an account of introspection is that many of these accounts are formulated assuming that introspection is exclusively first-personal. To argue for the necessity of IP by importing a conception of introspection that rests on an assumption of the necessarily first-personal nature of introspection would be question-begging.

Of course, appealing to an argument that introspection must be necessarily first-personal need not be question-begging. Shoemaker (1968) argues for the logical nature of IEM as a point against causal/perceptual theories of introspection. Suppose we identify the possessor of the mental state in a causal/perceptual way, by some sign concerning whose mental state it is – perhaps some kind of distinctive feeling. To use this to self-attribute a mental state, we would need to know that this sign signifies that the mental state is one’s own. But to know this, we would already need to know something of the form ‘I have mental states which involve this kind of feeling.’ This presumes that we can already identify ourselves in introspection, without noting the presence of such a sign. Shoemaker concludes that we can self-attribute mental states without this kind of identification, and, so he claims, without the possibility of misidentification.

Yet this part of Shoemaker’s argument establishes only this: when we make self-attributions through introspection, we do not require some contingently associated sign of ‘I’-ness. This does not yet show that introspection could not give us access to another’s mental states. It does not show us that we can’t misidentify another’s thoughts as our own through introspection, if we’re hooked up to their brain in some science-fictiony kind of way. It doesn’t even show us that introspection doesn’t operate causally. Note that we don’t identify ‘here’ by some distinctive perceptually accessible sign of here-ness (although of course we identify
features of our current location perceptually and through their closeness to our vantage point), yet we still see the world through a causal mechanism of perception.

Now, some of the reports of thought-insertion suggest that they are identifying thoughts belonging to the self by some sort of sign. For example, one patient noted of “inserted” thoughts that they in some way seemed her own “but I don’t get the feeling that it is...the feeling is not the same...the feeling is that it is somebody else’s.” (Hoerl 2001, 190). Yet one might identify something as belonging to another by a feeling without there being a different feeling by which you identify something as your own. A feeling of appropriately strong alienation might simply act as a defeating consideration for self-attribution; operation of such defeating considerations doesn’t require an appeal to lack of such defeaters when all goes well. So, there might be a feeling of non-agency or non-ownership, without requiring a corresponding feeling of agency or ownership.

Finally, there might be reason to think that introspection is not exclusively first-personal. Indeed, thought insertion delusions suggest that introspection is conceivably not exclusive in this way. If it is conceivable that introspection doesn’t deliver information exclusively about one’s own mental states, then perhaps this is possible. And if this is possible, then perhaps it is possible that one might introspectively detect another’s thought, but mistakenly self-attribute it. If we reject the conceivability of thought insertion, then we must think that there is some conceptual confusion in the experience of thought insertion delusions. I’m not sure that there isn’t, but I think we need more than the assumed necessity of IP to rest this on.

References
Byrne, A. 2005. Introspection. Philosophical Topics 33: 79-104
Campbell, J. 1999a. Schizophrenia, the space of reasons, and thinking as a motor process. The Monist 82, no. 4: 609–625.


---

iii Interestingly, in another paper (Campbell 1999b), Campbell argues that first-personal mental state ascriptions made on the basis of introspection have IEM merely because of the *contingent* fact that introspection delivers information only about one person. Campbell mentions thought-insertion in passing, but does not propose it here as a counterexample to IEM.

iv This seems to me to be the best way to think of the motor-process signaling that Campbell (1999a) discusses, if we accept Shoemaker's insights about identification: perhaps the signal that reaches conscious awareness in cases of thought insertion is a warning that something is going wrong, without this requiring an “all is going right” signal in the normal case.