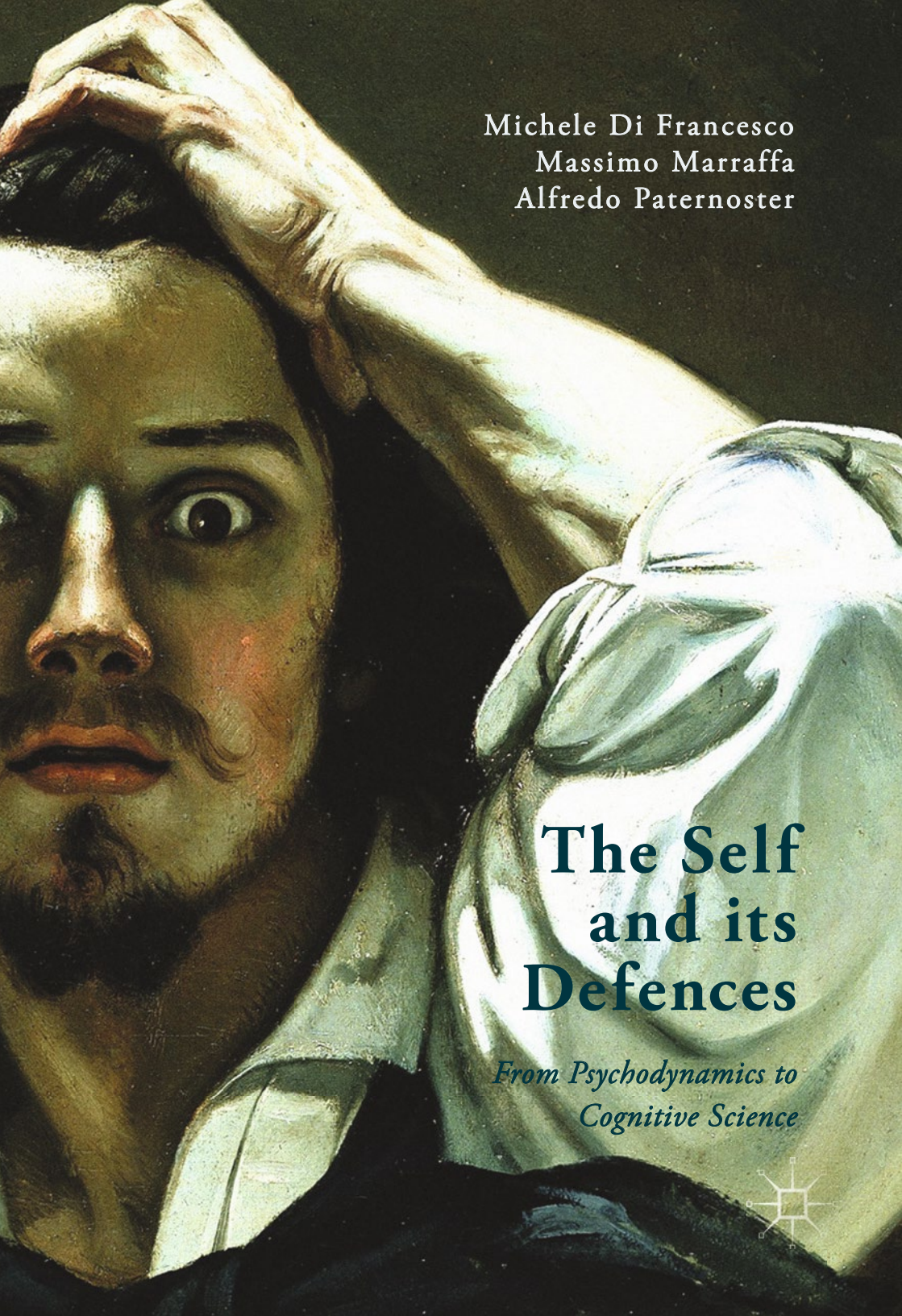

Michele Di Francesco
Massimo Marraffa
Alfredo Paternoster

The Self and its Defences

From Psychodynamics to Cognitive Science

The Self and its Defenses

Michele Di Francesco • Massimo Marraffa • Alfredo Paternoster

The Self and its Defenses

From Psychodynamics to Cognitive Science

palgrave
macmillan

Michele Di Francesco
School of Advanced Studies
IUSS Pavia, Italy

Alfredo Paternoster
University of Bergamo
Bergamo, Italy

Massimo Marraffa
University of Roma Tre
Rome, Italy

ISBN 978-1-137-57384-1 ISBN 978-1-137-57385-8 (eBook)
DOI 10.1057/978-1-137-57385-8

Library of Congress Control Number: 2016955530

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Macmillan Publishers Ltd.
The registered company address is: The Campus, 4 Crinan Street, London, N1 9XW, United Kingdom

Acknowledgements

This book is the result of a long-lasting cooperation between the authors. It arises from the confluence of two main lines of research. The first one relates to the themes of personal identity and the place of subjectivity in the world order. The second one pertains to the fruitfulness of a bottom-up, ontogenetic approach to human self-awareness, one that attempts to reconstruct how the complex psychological functions underlying the adult self-conscious mind evolve from more basic ones.

We have focused and pursued these lines of research for many years, and we obviously accumulated a great intellectual debt with quite a number of friends and colleagues, who offered criticism and advice, skepticism and support, over such a long period of gestation.

Many people created a lively intellectual environment where the seeds of many of the ideas expressed in this book could eventually breed. In particular, we are grateful to Luciano Arcuri, Grazia Attili, Lynne Baker, Sergio Fabio Berardini, Claudia Bianchi, Clotilde Calabi, Riccardo Chiaradonna, Roberto Cordeschi, Mario De Caro, Roberta de Monticelli, Rosaria Egidi, Carlo Gabbani, Rossella Guerini, Diego Marconi, Stefano Meacci, Mario Miegge, Simonetta Montanari, Roberto Mordacci, Michael Pauen, Giulia Piredda, Massimo Reichlin, Andrea Sereni, Alberto Voltolini. Michele Di Francesco owes a personal debt of gratitude to Stefano Cappa and Andrea Moro for sharing their knowledge and thus making the interaction between neuroscience and philosophy possible and indeed fruitful.

We were fortunate enough to be able to discuss the ideas developed in the book in workshops, conferences and seminars in various venues, such as Bucharest, Cracow, Granada, L'Aquila, London, Madrid, Milan, Pavia, Parma, Prague and Rome. We are grateful to the organizers and the participants—too many to mention by name—for numerous instructive comments and for their worthwhile advice.

George Graham's initial perplexities about the project of this book pushed us to a better understanding of what our real subject was. A book in Italian (Marraffa & Paternoster, 2013) and a series of papers (Di Francesco & Marraffa, 2013, 2014; Di Francesco, Marraffa & Paternoster, 2014; Marraffa, 2014, 2015; Marraffa & Paternoster, 2016) served as preparatory sketches of the final fresco we struggled to paint in this work. We express our warm gratitude to the reviewers who helped us in the process of clarification of our thinking. Philip Gerrans' valuable comments on Marraffa (2011a) allowed a better understanding of the theory of mind debate. Neil Campbell's suggestions on a draft of Marraffa and Paternoster (2016) addressed important revisions of the issues discussed in Chap. 3. Rupert Dörflinger's sympathetic and critical comments on Marraffa (2014) paved the way to the treatment of Locke's link between self-consciousness and responsibility in the Epilogue.

Special reference needs to be made to the influence exerted by the psychiatrist Giovanni Jervis on this book. Jervis was a prominent figure in the Italian intellectual landscape of the second half of the twentieth century. Marraffa, who was his student and friend, has argued in a series of papers (2011b, 2012, 2013) that Jervis delivers us the premises of a philosophical anthropology that aims to integrate Ernesto de Martino's phenomenological psychology of identity and the psychodynamic theme of defense mechanisms into the naturalistic framework of biological and psychological sciences. The fourth chapter of our book has grown from these powerful insights.

To Alfredo Tomasetta we are grateful for his thoughtful engagement with Chaps. 2 and 4, and for an endless series of discussions on (the limits of) naturalism. Only Cristina Meini read and commented on the whole manuscript; to her our debt is deep, since practically every chapter is better for her questions and suggestions. A special thank you is due also to Stefano Bacin for his (critical) comments on our reading of Kant.

References

Di Francesco, M., & Marraffa, M. (2014). A plea for a more dialectical relationship between personal and subpersonal levels of analysis. *Frontiers in Psychology, 5*, 1165.

Di Francesco, M., Marraffa, M., & Paternoster, A. (2014). Real selves? Subjectivity and the subpersonal mind. *Phenomenology and Mind, 7*, 118–133.

Marraffa, M. (2011a). Theory of mind. In J. Fieser (Ed.), *The internet encyclopedia of philosophy*. http://www.iep.utm.edu/theomind/

Marraffa, M. (2011b). Precariousness and bad faith. Jervis on the illusions of self-conscious subjectivity. *Iris, 3*(6), 171–187.

Marraffa, M. (2012). Remnants of psychoanalysis. Rethinking the psychodynamic approach to self-deception. *Humana.Mente, 20*, 223–243.

Marraffa, M. (2013). De Martino, Jervis, and the self-defensive nature of self-consciousness. *Paradigmi, 31*(2), 109–124.

Marraffa, M. (2014). The unconscious, self-consciousness, and responsibility. *Rivista internazionale di filosofia e psicologia, 5*, 207–220.

Marraffa, M. (2015). Mindreading and introspection. *Rivista internazionale di filosofia e psicologia, 6*, 249–260.

Marraffa, M., & Paternoster, A. (2013). *Sentirsi esistere*. Rome-Bari: Laterza.

Marraffa, M., & Paternoster, A. (2016). Disentangling the self. A naturalistic approach to narrative self-construction. *New Ideas in Psychology, 40*, 115–122.

Contents

1

Introduction: Setting the Stage

Reference to the notion of self plays a crucial role in a multitude of areas in philosophy and in social and human sciences; arguably most important, the notion of self seems to be an indispensable and central concept of the common-sense view of the world. It is the concept of an entity that, despite being extremely elusive and difficult to explicate, is the most fundamental piece of our mental life, something that makes all the rest of it possible. Despite this centrality, there is no consensus on what the self is, or even on its very existence.

In this book, we offer a theory of the self (which is at the same time a theory of self-consciousness, as will be clarified over the course of the book), whose core ideas are that (1) the self is a *process*, a psychobiological system activity of self-representing, and (2) this process aims mainly at defending the self-conscious subject against the threat of its metaphysical inconsistence. In other words, the self is essentially a repertoire of psychological maneuvers whose outcome is a self-representation aimed at coping with the fundamental fragility of the human subject. It is a *constructive* process that starts in the very early stages of our life and runs unceasingly throughout our entire life.

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_1

Our picture of the self differs from both the idealist and the eliminative approaches widely represented in contemporary discussion. Against the idealist approach, we deny that the self is something primitive and logically prior: a mental entity describable as the owner of its own mental states. Rather, we take it to be the result of a process of construction that starts with subpersonal unconscious processes. On the other hand, we also reject the anti-realistic, eliminative argument that, from the non-primary, derivative nature of the self, infers its status as an illusory by-product of real neurobiological events, devoid of any explanatory role. Our approach is then both *derivative* and *realistic*.

* * *

Our view of the self will be justified by a combination of philosophical arguments and data from cognitive sciences. The conceptual framework of our investigation can be described as *naturalistic*, *bottom-up*, and *systemic-relational*. Let us clarify each of these perspectives.

By 'naturalistic' we simply mean a framework that takes science seriously, at least in the sense that it is not possible for such a perspective to be in contrast with established findings provided by scientific disciplines. Even though we are not committed to taking our scientific view of the world as the only way to address the question of the self, we do consider recent findings in the realm of cognitive neuroscience and experimental psychology as a constraint upon it.

This brings us to the idea of a 'bottom-up' methodology. From Descartes' *cogito* to Husserl's transcendental ego, philosophy has adopted an inflationary approach to the self. One proceeds *top down*, starting from the philosopher's introspective self-consciousness, to arrive at everything else. The subject is taken to be transparent to oneself, and the knowledge provided by the reflective awareness that the mind has of its own structure and contents enjoys a special kind of certainty, which is distinct from our knowledge of the physical world. Our book invites the reader to take the opposite path: we start from the idea of the fruitfulness of a bottom-up, ontogenetic approach, which attempts to reconstruct how the complex psychological functions underlying the adult self-conscious mind evolve from more basic ones. This approach does not appeal to our introspective self-knowledge, but rather to the results of investigations into the

gradual construction of human self-awareness: from the automatic and pre-reflective processing of representations of objects (object-consciousness), through the awareness and then self-awareness of the body, up to introspective self-awareness and then narrative identity.

Our conceptual framework, however, aims to avoid not only a top-down ontologically inflationary approach to the self, but also an overly reductionist approach which explains *everything* in terms of bottom-up neurocognitive mechanisms. This is where a contextualist and systemic perspective comes into play. Here the individual's psychological problems are investigated by putting them in the inter-individual and social context in which they arise and obtain a sense. This systemic naturalism is rooted in the Chicago school of functionalism, and is the foundation of attachment theory—namely, the psychodynamic tradition within which we will develop our theory of self-consciousness.

The result of this multidimensional approach is a theory of self-consciousness according to which two aspects of the self are to be distinguished: on the one hand, there is a *selfing* process (the 'I', in Jamesian terminology), which is a synthesis function that works mainly at the subpersonal level; on the other hand there is the product of this process: the representation of the self (in James' words: the 'Me'), which is partly open to conscious inspection. The Me, which is constantly updated by the selfing process, is in the first place bodily, then psychological. The highest developmental point of this process is the narrative self, which is one among the layers of personality. This view involves a criticism of the primacy of self-conscious subjectivity, which, far from being a primary givenness, is unveiled as an articulate construction consisting of several neurocognitive and psychosocial components. As existentialist phenomenology puts it, we do not possess an essence that precedes our existence; our 'being-there' is always the being-there of a living body operating in a physical and social context, with a history. And it will be argued that this being-there is characterized primarily by its *precariousness*. In the absence of any metaphysical guarantee, the constructed self (the Me) is perpetually beset by the risk of its own disintegration. Hence the already-mentioned defensive nature of the self, its being primarily a process whose teleology is focused on self-protection or self-defence.

As the reader can already realize from these introductory remarks, there are several strands in this book. In particular, it combines cognitive psychology, analytical philosophy and psychodynamics (not to mention some

excursions into a 'continental' philosophical anthropology). In a vague but (we hope) understandable sense, the result is more an exercise in the philosophy of psychology than in the metaphysics of mind—even if our naturalistic methodology renders the boundaries between epistemological and metaphysical worries somewhat vague and undefined. Nor do we propose a systematic comparison with the classical phenomenological approach. We simply follow our route from subpersonal unconscious processes to the personal conscious self-representation and in doing so we address metaphysical or phenomenological problems as they present themselves.

* * *

Let us now give an overview of the structure of the book.

The second chapter is devoted to an analysis of the notion of unconscious, both in the cognitivist sense (the so-called 'cognitive unconscious') and in the Freudian sense. We explain why cognitive sciences focus on unconscious processes and structures, strongly diminishing the importance of the conscious level, and we determine what is alive and what is dead in Freud's theory of the unconscious. Starting from this analysis, we argue that the strategy, pursued in cognitive science, of explaining behavior and mental phenomena with unconscious or *subpersonal* processes and structures is fruitful. However, since this approach runs the risk of overextending the scope of the concept of mind (this is the 'mark of the mental' problem), and of making the problem of unifying personal-level explanations with scientific explanations of mental phenomena (the 'interface problem') more difficult, we also make a case for a dialectical relationship between personal and subpersonal levels of analysis. In particular, we submit that certain psychodynamic constructs very close to the personal level (paradigmatically, the notion of attachment) are indispensable to an account of self-consciousness. The chapter thus ends with the development of the psychodynamic framework within which to conduct our research on self-consciousness. We focus on relational themes, especially on the forms of cognitive-affectional relationality of the very young child. As is shown by the theories of object relations and attachment, physical contact and the construction of protective and communicative interpersonal structures constitute the infant's primordial psychological needs, around which her mental life gradually takes form.

In the third chapter, we undertake our realist (neither idealist nor eliminative) view of the self, arguing that the first and fundamental form of self-consciousness is the consciousness of one's own body, taken as a whole. We start with a criticism of the 'exclusion thesis', the claim that there is no room for something like the self in the natural order—a thesis that in modern philosophy goes back to Hume's and Kant's criticisms of the Cartesian self. After having dismissed the Humean eliminative approaches to the self, we turn to a critical examination of two different approaches to the theme of self-consciousness. The first perspective is that of analytic Kantianism, a line of thought that stems from Peter Strawson's *The Bounds of Sense*; the second perspective is the project to provide a naturalistic version of the phenomenological claim that conscious experience entails self-consciousness, which has been pursued especially by Dan Zahavi.

The trouble with the former, whatever its intrinsic merits, is that it is unable to provide a genuinely *empirical* account of self-consciousness: the Kantian tradition is a form of a priori philosophical psychology, or, better, transcendental epistemology, which, insofar as it is empirically unconstrained, is incompatible with our naturalistic approach. Instead of a transcendental synthesis, we posit a *psychobiological* synthetic function: the already mentioned *selfing* process. Moreover, and as a consequence of its purely conceptual character, Kant's theory of self-consciousness hinges on a view of the human subject as originally unitary; we argue, in contrast, that the subject is primarily non-unitary and gains a sense of unity in the act of raising a bulwark against the threat of not being there.

Against the phenomenological project, we show that there is no pre-reflective or non-reflective self-consciousness that accompanies every conscious state from birth. This is an empirically void construction, ultimately still reminiscent of Kantian transcendentalism. The outcome of this discussion is that the most minimal form of self-consciousness is *bodily* self-consciousness, the capacity to construct an analogical and imagistic representation of one's own body as an entire object, simultaneously taking this representation as a subject, that is, as an active source of the representation of itself. In the last section of this chapter we begin to outline, building on James, our 'processual' view of the self: we distinguish between the self as the interminable objectivation process (the I) and the self as the multidimensional representation continuously updated by this process (the Me).

Chapter 4 is devoted to the development of the *psychological* self: an account is given of how the awareness of ourselves as subjects who are bearers of mental states is constructed from the awareness of one's own body. We show that our inner world evolves through an interplay—modulated by sociocultural variables—of mentalizing abilities, autobiographical memory and socio-communicative skills. The starting point is a critical discussion of introspection: following Freud's idea that our inner life is saturated with self-deception and bad faith, we show, based on the enormous amount of confabulation data from cognitive neuropsychology and social psychology, that our knowledge of our mental states is to a large extent inaccurate. Far from realizing that our actions are actually determined by unconscious mechanisms, we 'fabricate' rational post-hoc explanations of our behavior by means of an incomplete, partial and, in many cases, seriously defective folk theory of psychology. Thus, where Descartes saw a given essence (the self-transparent consciousness-substance), there is now something *constructed*, the product of an apparatus that allows us to partially describe, and above all narratively justify, fundamentally unconscious mental processes. With this result in hand, we focus on the ontogenesis of the inner, virtual 'theater' of the mind, arguing that the construction of an introspective experiential space occurs through the process of turning one's mentalistic skills—the ability to ascribe mental states to others—upon oneself under the communicative pressure of micro-social contexts. We will look firstly at *affective* mentalization, arguing that a positive attunement in proto-conversational infant-caregiver interactions plays a crucial causal role in the construction of the phenomenology of basic emotions. We will then examine how the construction of an inner experiential space advances under the thrust of caregivers' mind-minded talk. Finally, we turn to the most mature and cognitively demanding stage in psychological awareness, that is, the development of a narrative or autobiographical self. Here we highlight the importance of the sociocultural context: data from cultural psychology show that psychological self-consciousness is not an all-or-none phenomenon; the incompleteness of the capacity to conceptualize the existence of an inner experiential space has been observed in normal adults in pre-agricultural or pre-literate agricultural cultures.

In the fifth chapter, we put forward our central thesis about the nature of the self: the idea that the self is essentially a collection of defensive strategies aimed at coping with its lack of a metaphysical guarantee. Indeed, psychological self-consciousness, far from being a stable faculty, is a precarious acquisition, continuously under construction by the subject and constantly exposed to the risk of crisis. This precariousness is, therefore, the key to grasping the defensive nature of narrative identity. Defensiveness is immanent to human self-consciousness, since the latter constitutes itself precisely in the act of taking measures against its own dissolution. The chapter concludes with a clarification of the difference between our position and eliminative accounts (such as Dennett's) about the self. We show how our naturalistic approach to the narrative self also enables us to reject the antirealist argument that infers, from the non-primary, derivative nature of the self, a view of it as an epiphenomenal by-product of neurobiological events or, alternatively, of social (or socio-linguistic) practices. The antirealists—we will argue—disregard the essential psychodynamic component of identity self-construction. The need to construct and protect the most valid identity possible is rooted in the subject's primary need to subsist subjectively, and thus to exist solidly as a describable ego, as a unitary subject. Far from being the staging of an ephemeral self-deception, the incessant construction and reconstruction of an acceptable and adaptively functioning identity is the process that puts into place our intra and interpersonal balances, and is thus the ground of psychological well-being and mental health. Unlike Dennett's Joycean monologue, in our model self-narrative is not mere empty chatter: it is a causal center of gravity. In this sense, the psychodynamic component of our theory plays a crucial role in shaping our 'robust' (i.e., genuinely realist) view of the self.

2

The Unconscious Mind

In the last 50 years, the sciences of the mind have been mostly concerned with unconscious functions. Indeed, the mental processes studied by cognitive science, such as perception, reasoning or language understanding, are not accessible to consciousness. Only their inputs and outputs (and perhaps some fragmentary parts) are. We are aware of the final results of the processes, but not of their internal dynamics. In this perspective, the unconscious is, in a way, far more important than the conscious, insofar as it is the unconscious which *explains* the abilities manifested in our behavior.

On the other hand, this emphasis on 'hidden' processes resulted in our losing what we are inclined to regard as the mental *par excellence*: the contents of our flow of consciousness, the phantasmagoric pattern of sensations and emotions which constitute our mental life—that is, losing our self-conscious subjectivity. But, if one is not talking about *conscious* mind, is one really talking about mind at all? The answer to this question depends crucially on what one takes the *mental* to be. This is the so-called issue of the *mark of the mental* (or *the cognitive*), recently brought to prominence in the debate on the extended mind (the view according

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_2

to which cognitive systems go beyond the boundary of the organism). After discussing the criticisms leveled by John Searle against the notion of unconscious mind, we propose a notion of the mental which is able, on the one hand, to account for the relevance of unconscious functions to understanding our mental abilities, and, on the other hand, to accord to consciousness an important role in the characterization of the mental domain. In this way, we avoid the risk that, in cognitive science, the concept of the self and the related concept of consciousness end up constituting a somewhat bothersome remnant.

Our clarification of the conscious/unconscious distinction will enable us to develop the psychodynamic framework in which to conduct our research on self-consciousness. After determining what is alive and what is dead in Freud's theory of the unconscious, we will examine a central characteristic of the development of post-Freudian psychoanalysis, namely, the focus on relational themes—especially on the forms of cognitive-affectional relationality of the very young child. The development of the theories of object relations and attachment is part of this trend: here, as we will see, physical contact and the construction of protective and communicative interpersonal structures constitute the infant's primordial psychological needs, around which her mental life gradually takes form.

2.1 The Mind and Cognitive Science

Our scientific knowledge of mental phenomena is today provided by *cognitive science*, a collection of disciplines that aim to explain how we are able to perceive, reason, understand language, make rational choices, plan and perform actions; in brief, all the capacities that are considered as distinctively mental.[1] We could say that cognitive science aspires to investigate human nature across the board. This ambitious goal reveals a crucial but controversial presupposition, that is, that *human* nature, and specifically the mind, is indeed a natural fact, and as such constrained

[1] It could be argued that talking about cognitive science*s*, at plural, is more appropriate. Much depends on the importance one accords to the differences between the research programs in the field. We will not be concerned with this problem here, and we will freely use the singular and the plural form without being committed to a certain epistemological position.

by the biological laws of our species. Of course, not all human behavior rests on biology, but the challenge of cognitive science—given its vocation and method—consists exactly in widening, as far as it is possible, the naturalistic realm, in denying that our choices and actions can be exhaustively attributed to historical and interpretative factors, and thus trying to overcome the dichotomy, dear to the hermeneutic tradition, between *Naturwissenschaft* and *Geisteswissenschaft*.[2]

A distinctive feature of the development of cognitive science has been the continuous and significant growth of importance of neurosciences. While in the 1960s and 1970s findings about the brain played a negligible role in explaining mental capacities, they currently occupy a central position. That many readers have likely heard about neuroethics, neuroaesthetics, neuropolitics and even neurotheology is a telling indication of an 'outbreak' of inquiries into the brain that have led some researchers to suspect that old ideas may have been presented as novel just by changing the word 'mind' to 'brain', without bringing about any actual scientific progress. In fact, the beneficial circumstance that there exists an interplay, and, within some limits, integration, between psychology and neuroscience, does not change the conviction of several researchers that the respective subjects are very well distinguished from each other. After all, this is also the folk intuition, according to which the relation between mind and brain, however close, cannot simply be couched in terms of an identity. If we can easily distinguish brains from persons—in the common sense view the brain remains, in spite of all its extraordinary importance, a physical organ, on a par with the heart or the stomach, we have difficulty in finding a firm collocation for the mind, which cannot be identified with either a person or the brain, despite being closely linked to both. Of course, common sense is not necessarily our pole star; and here, as in other cases, science must to some extent distance itself from it. We shall have to see how, where and to what extent.

* * *

[2] One of the shared assumptions in cognitive science is that, although human beings have capacities that animals do not have (but note that the opposite is also true), first of all language, there is no radical discontinuity between human and non-human natures. Of course, this is Darwin's lesson.

Much of modern common sense about the mind comes from Descartes. According to Descartes's theory of mind, the mental dimension coincides with the conscious dimension. The mind is *res cogitans*, and thought, its defining attribute, is explicated in terms of awareness (*conscientia*). The Cartesian thesis of equating the mental dimension with whatever lay within the scope of one's consciousness is partly endorsed by common sense. 'Partly' because, under the influence of psychoanalysis, today's common sense view of the mind has incorporated the idea of unconscious mental states. However, as we see in Sect. 2.2, in this folk culture of the mind there still prevails the most evident limitation of the Freudian view of the unconscious: the unconscious is but a different kind of conscious mind, in the sense that it has a structure similar to the conscious and has the possibility to become conscious.

Moreover, in Descartes's model of the subject, the mental dimension is radically distinct from the body. The body is bound by mechanical laws, is located in space, and is decomposable into parts; by contrast, the mind is free and creative, with no spatial location, and is an indissoluble unity. The Cartesian conscious mind is the locus of personal identity and (as shown by the *cogito* argument) could persist even if the body and the external world were illusory. In other words, this idea of mind is the result of a secularization of the idea of soul: like the soul, the mind is viewed as an essence that precedes existence, namely, a set of spiritual prerogatives that are *primary*, and hence *essential*, in comparison with the *accidental* nature of people's bodily determinations.[3]

In the twentieth-century sciences of the mind, however, both Cartesian assumptions have been rejected. The distinction between mind and body is widely denied on ontological grounds because any mental process depends on the brain and is *realized* by the brain: human beings are evolved biochemical machines.[4] In addition, the mental dimension is no longer

[3] Edward B. Tylor (2010) was the first to formulate the hypothesis that the natural tendency to a spiritualistic objectification of the mind (and hence the idea of soul) is due to spontaneous rationalizing mechanisms. Later, the idea of the natural origin of dualistic thinking was pursued by other thinkers, most notably by Jean Piaget. More recently, the hypothesis has been suggested again in psychology (Barrett, 2004; Bloom, 2004; Boyer, 2001) and in anthropology (Astuti, 2001; Cohen & Barrett, 2008).

[4] Obviously in the philosophical debate dualistic positions are still present—even if very rarely they take the form of Cartesian substantial dualism. In any case the 'mainstream' ontology of cognitive science radically denies immaterial entities.

confined to the conscious one since most of the phenomena and processes that are taken as mental by cognitive sciences are not conscious. In order to fully understand this overturning of the Cartesian approach, with special regard to the dissociation between mind and consciousness, we need to dwell on two epistemological assumptions that played a crucial role in the development of the current sciences of the mind: (1) the idea (ascribable to Alan Turing) that mental processes have a *computational* nature; (2) the idea (ascribable to Noam Chomsky) that intelligent behavior is mediated by *mental representations*. These two assumptions, together with skepticism about introspection, imply the claim at the core of our discussion in this chapter: *the dissociation between mind and consciousness*.

2.1.1 The Computational-Representational Mind

Historically, one of the most important arguments against the possibility of conciliating the ordinary and the scientific images of the mind is the idea that no merely mechanical system could ever show genuine intelligence. No wholly mechanical system—the idea runs—can show flexible, open-ended, creative intelligence: do something truly new, respond intelligently to the unexpected. But humans do have such capacities! So, for much of the history of modern philosophy, the prevailing wisdom was that the human mind is not merely a complex physical mechanism of some kind (though animal minds might be).

Over the last 100 years or so, however, this view has been increasingly seen as untenable. In particular, Alan Turing's seminal work on the mechanization of intelligence seems to refute the above-mentioned argument for irreconcilability. For this work—and the computer science and AI that have flowed from it—seems to show that even activities that we would consider as intelligent and creative, such as reasoning or language understanding, are within range of a machine, of a purely mechanical device. The leading idea is that a particular intelligent task can be accomplished mechanically if it is decomposed into a sequence of elementary steps, each of which is well-defined, completely specified (without ambiguity) and sufficiently basic to be readily carried out by any 'executor' whatsoever.

Think of a cooking recipe for dummies, where nothing is taken for granted, including even such an obvious instruction as 'turn on the stove' or 'put the pan on the fire'.

These kinds of procedures are termed *computations* (or *algorithms*). More precisely, the concept of computation is the logical-mathematical formalization of the intuitive concept of procedure. But for our purposes, we need not be too rigorous. Just think of a computation as a computer program—a connection anyone even slightly familiar with computer science will already have made. Computers are able to perform in an intelligent way because they are programmed machines; in other terms, any intelligent activity can be accomplished through a proper sequence of basic operations: the 'right' program. So the idea is that mental processes can be characterized as computations, as computer programs.[5]

Computations, of course, work on data. So, the hypothesis that mental processes are computations requires that different kinds of information (visual, auditory, linguistic, etc.) be encoded or *represented* (see below) 'in our head' in some format suitable to their being processed. Like a computer program, a mental process processes input information and outputs other information. Of course, this is not to claim that the brain literally works like a computer—which is simply false—but that the processes realizing our cognitive capacities can, *at a certain level of abstraction*, fruitfully be modeled on computational processes. We will examine this point further in Sect. 2.1.2.

Therefore, saying that a mental process is a kind of computation is the same as saying that it is an information-processing process. The data on which computer programs operate need not be numerical: they can concern any domain of knowledge provided that the relevant information is *encoded*, that is, expressed in a description that can be understood by the executing system, for example in some programming language. Likewise, in the case of the mind, information concerning our bodies and the world around us must somehow be encoded in order to be processed

[5] The conception of the mind as a computational device was already put forward by Hobbes and Leibniz. But only with Turing did this intuition become a sound, grounded hypothesis able to foster a serious research program.

by the computational processes executed by the brain. The processes of visual perception, for example, must process information about the shape, color, distance and possible movement of an object in the visual field. The processes of language understanding must process information concerning linguistic sounds, syntactical structure, and literal and intended meanings of sentences. Thus mental processes manipulate—build and modify—informational structures that can be characterized as (mental) *representations* inasmuch as they are entities that stand for objects in and properties of the world. Just as a picture or a linguistic expression represents an object in the external world, say a red rose, conveying information about it, there are structures in our heads that represent objects and properties of the world.

We have introduced the idea of mental representation starting from the observation that a computational process needs to work on some data or pieces of information, but the concept of representation must also be considered under another, more significant aspect, from the viewpoint of the history of psychology; the concept of mental representation marked the transition from behaviorism to cognitivism.

In the late nineteenth and early twentieth centuries, scientific psychology was predominantly a psychology of introspective consciousness. Pursuing the project to make introspection a rigorous method of inquiry, and thereby to upgrade psychology to the status of the other natural sciences, early experimental psychologists meticulously probed the contents of consciousness in an effort to offer a full description of the mental landscape as it appears to the subject. In short, this psychology was a kind of phenomenological investigation of self-aware subjectivity.

By virtue of the mentalistic idiom, these introspectionist psychologists experienced no reluctance in talking to 'poets, critics, historians, economists, and indeed with their own grandmothers. The nonspecialist reader in 1910 would be in equally familiar territory in William James's *Principles of Psychology* and in the novels of James's brother Henry' (Stich, 1983, p. 1). John Watson's brand of behaviorism put an end to the friendly relations between scientific psychology and folk psychology, urging the abandonment of the introspectionist attempts to make consciousness a subject of experimental investigation. Still more radically: since 'mind' in the folk sense of the term refers to something mysterious and unfathomable, the mind was

to be banned from scientific discourse, or at least re-conceptualized as the organism's mere potentiality or disposition to generate behavior. A psychology aspiring to scientific respectability had to rely only on publicly observable data, that is, patterns of responses (overt behavior) to stimuli (physical events in the environment). The outcome was an extremely austere conception of psychological explanation: the psychologist, equipped with nothing but Pavlov's conditioning and Thorndike's law of effect (a precursor of Skinner's operant conditioning), had to explain why we behave as we do, without making appeal in the explanation to unobservable theoretical entities like ideas, imagery, intentions and so on. What occurs in the 'head', between input and output, was a topic for physiology (the ultimate behavioral science). An organism, as the behaviorist viewed it, was 'empty'.[6]

After the 1930s and 1940s, an increasing perception of the limits of the S(timulus)-R(esponse) explanation led behaviorism to evolve toward what would become, starting from the 1960s, cognitive psychology. A landmark in this evolution was the classic series of rat experiments in the Berkeley laboratory of Edward C. Tolman (1948). These experiments showed that the maze-navigation behavior of rats could not be explained in terms of S-R mechanisms, suggesting that the animals were building up complex representational states, or 'cognitive maps', which helped them locate reinforcers. Tolman's conclusion pointed in the same direction as the hypothesis that Kenneth Craik had put forward five years before in *The Nature of Explanation* (1943): the mind does not work directly on reality, but rather on 'small-scale models' of it.

The time was ripe for psychology to resume dealing with what is in the head: in order to account for behavior, psychology must be a science of the mental structures mediating it, that it is to say, a science of the mind. *Mental representations*, rather than stimulus-conditioned responses, are able to account for behavior. Having mental representations enhances an organism's capacity, making its behavior more flexible, because when some features of the environment are not present or manifest, they can (at least in some cases) be represented: 'something else can stand in for them, with the power to guide behavior in their stead' (Haugeland, 1998, p. 172).

[6] 'Empty organism' is the term used by E.G. Boring to characterize Skinner's position (cited in Newell & Simon, 1972, p. 875).

The claim that behavior is driven by representations was the decisive move beyond the narrow limits of behaviorism, paving the way for a no less rigorous study of what is *inside* the head. And this move was epistemologically justified, since postulating *unobservables,* such as electrons and genes, is standard practice in science.

Some ingenious attempts to refine the S-R schema were made to account for Tolman's experimental results without his troublesome mentalistic concessions (see Hull, 1943). However, such a schema turned out to be totally powerless when the focus shifted from maze-navigation behavior in rats to verbal behavior in human beings. Thus, it is hardly surprising that one of the main factors of the transition from behaviorism to cognitivism was the rapid development, beginning in the late 1950s, of a mentalistic theory of language, namely Noam Chomsky's generative linguistics.

Chomsky introduced the use of the term 'representation' in cognitive science with reference to the rules of natural language grammar. According to the groundbreaking linguist we master a language's grammar because its rules are recorded (i.e., represented) in the head. Grammar rules establish which strings of words form a sentence and which do not: a grammatical or *well-formed* sentence is a sentence that is generated in compliance with rules; a pseudo-sentence, which is to say an ungrammatical sentence, cannot be generated on the basis of the rules. For instance, there are no rules (in English) that can lead a speaker/hearer to produce or accept 'with the runs John dog'.

The precise form of these rules is not important here. However, there are two points to be highlighted: (1) positing rules is *needed* to account for facts about linguistic behavior—in this case the capacity, already mastered by children around three years of age, to produce and recognize grammatical sentences; (2) the rules are 'inscribed' into the computational structure, to the extent that they are, as we usually say, 'hardwired' in the brain; this means that linguistic processes work in accordance with these rules and we cannot but follow them, in the same way that we cannot prevent ourselves from seeing the wardrobe before us when opening our eyes upon waking. Indeed, since we have no awareness of these rules, we could not decide whether to follow them or not (see below). Something similar applies to other cognitive processes, such as perception or reasoning; they follow different rules or representations, but they are still driven by rules and representations.

To conclude, the view of the mind as a collection of computational (or information processing) processes can seamlessly be combined with that of the mind as a representational system. The informational structures that are manipulated by 'mental programs' are representations to the extent that they convey and encode information about the environment. Reference to a mental representation is an *explanatory hypothesis* within a theory of cognition as information processing: a representation is something that a mental process described in terms of information processing has to build—compute—in order to give rise to a specific behavior (Cummins, 1997). In some cases, such as that of grammatical rules, there are good reasons to think that representations are innate, already embedded in the system.

2.1.2 The Dissociation Between Mind and Consciousness

It is not difficult to see how the new conception of mind delineated in the preceding section breaks the close association between mind and consciousness. The mental processes investigated by cognitive science, as well as the mental representations that it posits, *are not conscious*.

Let us consider, for example, the case of language. Our understanding of a sentence is immediate. We instantly know whether or not we have grasped (as usually happens) what our interlocutor is telling us. And yet a lot of machinery is needed to understand a sentence: a nearly continuous sequence of sounds must be segmented into words, that is, into meaningful units; a grammatical structure must be associated to the sentence, and this structure is not always the only one possible (hence, one needs to choose the right one); ambiguous or polysemic words are to be interpreted in a manner appropriate to the context, etcetera. We have no awareness of all these complicated processes, just as we have no awareness of the structures of information—the representations—that must be built up to successfully perform these tasks. We are not conscious of having grammar rules inside our heads and of systematically applying them during the processes of understanding.

Another example: visual perception, one of the most successful research areas in cognitive sciences. A computational theory of vision aims to answer the question 'How do we see?' 'To see' is a verb that refers to our ordinary visual experience; for example, I lie down on the sofa with my eyes shut; I open my eyes wide and can say that I see a table and a chair. But, as in the case of language, our brain has to perform a multitude of operations to achieve this apparently simple result. Visual experience is the final outcome of an extraordinarily complex process that begins with the photoreceptors in the retina transmitting, along the optic nerve, electric signals that encode the level of light energy absorbed by the photoreceptive cells; later, various stages of processing occur, realized by different brain circuits which assemble the various pieces of information concerning form, color, movement, etcetera, into increasingly complex structures, and finally integrate them into a single coherent 'percept'. Marr's theory, the standard model of the computational theories of visual perception, assumes three stages of processing, in each of which specific representation of visual features is constructed; for example, the earliest stage of processing builds a representation, called a 'primal sketch', in which the strongest discontinuities in light intensity (the so-called 'zero-crossings') on the retinal image are detected. This clue is of great importance for the visual system since such discontinuities are very likely to correspond to the contours of an object.

Now, here, as in the case of the understanding of a sentence, we have no awareness of the unfolding of the processes underlying the construction of the percept. We do not notice what occurs in the eye and in the brain. What we are aware of is the final outcome, and the final outcome is an admirably harmonious and integrated world that presents, or better, imposes itself upon our consciousness. All that we can knowingly do is to shift our visual attention: deciding where to look, moving so that we can access parts of the world that earlier were outside of our visual field (e.g., the back of the object in front of us).

* * *

Thus, although the *explananda* of cognitive science are often conscious phenomena (understanding a sentence, having a visual experience, drawing the right conclusion from two or more premises, etc.), their *explanans*—processes and representations—is situated at a completely unconscious

level. The task of cognitive psychology and more in general of cognitive sciences consists precisely in bringing to light these unconscious mechanisms. Conscious phenomena are but episodic fragments of an incessant cerebral activity. Moreover, the fact that some representations can emerge into consciousness, as happens, for example, with the form or geometry of the visible surfaces of an object, is not relevant for the theory: the role that a representation plays in a cognitive process does *not* depend on whether it is conscious or not.

It is important to point out that the unconscious character of the mental processes investigated by the sciences of the mind remains as such through the development of this research program. Many researchers no longer acknowledge—or at least do not completely acknowledge—the computer-inspired model of the mind, and the years since the 1980s have witnessed a dramatic increase in the importance attributed, on the one hand, to the role of the brain and, on the other, to that of the environment.[7] Nevertheless, consciousness continues to play a secondary role in the explanation of the working of the mind. A single example will suffice: according to a theory currently enjoying increasing success in cognitive neuroscience, understanding action-related sentences (e.g., 'John ran' or 'Mary firmly grasped the handle') involves the activation of pre-motor areas, those in which mirror neurons are found.[8] This phenomenon has been interpreted as unconscious simulative activity: in order to understand the sentence, people simulate in their *unconscious* minds the execution of the action to which the sentence refers. It is only a *simulation* because the relevant action is not actually executed. In a sense, it is an imaginative process—when hearing 'Mary firmly grasped the handle' one *imagines* grasping a handle, without performing any movement—which, however, does not surface into consciousness: it is definitely not necessary to consciously imagine grasping something in order to understand the verb 'to grasp'. The use of the verb 'to imagine', however, seems inappropriate, since in common usage imagination is a conscious activity, and it is perhaps also for this reason that the term 'simulation' has been preferred (Paternoster, 2010).

[7] These are the so-called 'vertical' and 'horizontal' expansions of cognitive science. See Bechtel, Abrahamsen, & Graham (1998), p. 77.

[8] Mirror neurons are nerve cells that 'fire' both when their 'owner' is performing some action (e.g., grasping an object), and when she sees someone else performing the same action.

It can be clearly seen, then, how even the object investigated by the 'new' cognitive sciences is still constituted by non-conscious processes, and how these are considered as genuinely *mental*, despite the predominant role that neuroimaging data plays in these studies. And it can be clearly seen how the phenomenological data—the content of our consciousness—are largely irrelevant for the theory.

2.1.3 Levels of Explanation

The computational-representational view and the related emphasis on the unconscious, that is, on *subpersonal* processes and representational structures, raise at least two problems. The first concerns the nature of the relation between this kind of explanation and the personal level explanation, that is, the ordinary view of the mind. The second problem concerns the meaningfulness itself of regarding subpersonal computations and representations as genuine pieces of the mind. This is an assumption that, far from being obvious, requires justification. The two problems are connected, since if consciousness is a marginal and unnecessary ingredient of mind, then linking ordinary psychological explanation to scientific explanation turns out to be quite difficult: scientific psychology seems to have nothing to say on the topic of persons and about what mind is for us.

Section 2.3 is devoted to the issue of the legitimacy of regarding the subpersonal dimension as genuinely mental. Let us now spend some words on the problem of the relation between ordinary explanations and scientific explanations, which has been happily labeled 'the interface problem' (Bermúdez, 2005).

In its more general form, the interface problem consists in the difficult task of showing how explanations expressed in a folk mental vocabulary (including terms such as 'belief', 'desire', 'intention', etc.) could be linked to a variety of scientific explanations, whatever their form: computational, neuronal, etcetera. Note that the computational-representational view is already an attempt to address the interface problem, to the extent that it more closely associates mental facts to brain facts: computational-representational states are neither properties of persons nor properties of brains, and this fits well with the intuition that, as we said at the

beginning, the 'location' of the mind is somewhere in between persons and brains. In other words, projecting ordinary mental states (which are properties of persons) onto computational-representational states seems to be less difficult than projecting ordinary mental states directly onto brain states.

This strategy could be expressed in terms of a collection of explanatory layers or levels. At the highest level, there is the common-sense mental explanation, namely, folk psychology. The core of folk psychology is the idea that behavior is causally explained by mental states such as intentions and beliefs (e.g., 'I went to Perry's bar because I wanted to talk with Clare, and I thought she would be there'). Folk psychology is the ordinary image of ourselves as mindful persons. Freudian psychoanalysis can also be included in folk psychology—or, better, it is an *extension* of folk psychology—to the extent that it accepts its theoretical entities, such as desires and beliefs, as well as their causal role for action (see Sect. 2.2).

Descending to the next level, we find computational explanations. As mentioned above in some detail, these represent a sort of scientific psychology, which postulates certain *subpersonal* or unconscious mental entities and takes them as the causal factors responsible for behavior.

Finally, one step further down, there are neuronal explanations, which account for upper level facts in terms of cerebral facts (and we could, in principle, go further down to the level of molecular explanations, but we are not interested in this aspect here, since this step, far from especially concerning the mental/cerebral domain, can take place in any case of biological explanation).

This multilayered model of explanation has seemed to many a good idea—and we agree. Yet the interface problem persists: what is the relation between folk psychology and the underlying (computational) psychological explanation? In particular, what are the relations between the entities postulated at the higher level (such as beliefs and desires) and entities postulated at the lower level (the subpersonal structures postulated by scientific psychology)? Straightforwardly identifying a personal mental state (e.g., a belief) with a given subpersonal structure could be tempting, but the reality is far more complicated.

A good way of framing the question of the relation between personal and subpersonal explanations is in terms of a tension between dependency and

autonomy. On the one hand, mentalistic conscious explanations appear to be *dependent* on lower-level explanations. Cognitive sciences have shown *ad abundantiam* that what happens at subpersonal levels determines or at least affects what happens or seems to happen at the personal level. On the other hand, common sense explanations of behavior appear to be effective in many cases, and able to single out those processes that are causally relevant for the genesis of behavior in a variety of social contexts, such that they appear to be hardly dispensable. These explanations, however, appear to be *autonomous* for two reasons: first, they do not seem to require any reference to further facts (external to them); second, they make use of principles and explanatory styles that are different from those brought to bear in subpersonal causal explanations. Indeed, they embody principles of rationality, holistic approaches and reference to hardly naturalizable notions such as content or action.

Thus, the distinctive difficulty of the interface problem can be couched in the form of the following dilemma: either one takes seriously the lay conception of the mind (but this way the unification with lower layers becomes very unlikely), or one does not take seriously the ordinary view of mind (but in this way we are not able to account for some widely shared and strong intuitions; first of all the idea that folk mental states are causes of behavior).

Now, the approach of many philosophers who, like us, are inclined to take seriously the development of cognitive science, consists in trying to solve the interface problem by asking how and to what extent the common sense conceptual picture should be modified on the basis of the results of cognitive science. Note that, put this way, the closeness to folk psychology no longer represents an asset of psychoanalytic theory. For, to anticipate a point that we will consider more fully in the next section, the unconscious that actually reveals the causal processes underlying behavior is the computational unconscious, whereas the psychoanalytic unconscious is not able to realize this ambition. In fact, although psychoanalysis starts from the level of the person to reach the subpersonal level, it 'ends up having to re-import the personal level at the subpersonal, in order to get all the subpersonal bits to do what they are supposed to do' (Gardner, 2000, p. 100). In this sense, psychoanalysis is a failed attempt to give up the personal for the subpersonal.

* * *

In summary, cognitive sciences aim to study the processes implementing the capacities underlying intelligent behavior (perceiving, reasoning, understanding language, etc.). Such processes are inner and broadly unconscious. They are, as philosophers of mind usually say, *subpersonal* insofar as, not emerging at the level of consciousness, they are not 'owned' by the person, who cannot access them. In spite of this, they are considered as genuinely *mental*. Consequently, consciousness is not the constitutive essence of the mental. Clearly, this clashes with the ordinary view of mind, making it very difficult to link scientific explanation to personal-level explanation. Therefore, a justification is required.

2.2 The Freudian Unconscious

We assigned two 'founding fathers', Turing and Chomsky, to the basic assumptions of cognitive science—the mind as the processor, and as the representational system. If we want to assign one to the claim of the dissociation between mind and consciousness, the name of Sigmund Freud immediately springs to mind. According to Jerry Fodor, for example, it is to Freud's credit that he challenged the supposedly inextricable link between consciousness and intentionality: 'He made it seem plausible that explaining behavior might require the postulation of intentional but unconscious states. Over the last century, and most especially in Chomskian linguistics and in cognitive psychology, Freud's idea appears to have been amply vindicated' (Fodor, 1991, p. 12).

However, this historical note needs to be rectified. Cognitive sciences have not simply vindicated Freud but have gone much further. For in cognitive sciences unconscious phenomena are something radically different from the Freudian unconscious. This distinction between the cognitive unconscious and the Freudian unconscious is now our focus.

* * *

According to Descartes, we have a transparent awareness of our own mental processes and contents: 'there can be nothing within me of which I am not in some way aware', he writes in the first replies to the objections raised against his *Meditations* (1984, p. 77). There is no room for the notion of unconscious mentality here.

During the second half of the nineteenth century, however, the unconscious insistently claimed its own rights. Neurologists and psychiatrists drew attention to phenomena such as convulsive great hysteria, dissociative fugue, or multiple personality disorder, which could hardly be reconciled with the consciousness-dependent conception of mind originating from Descartes. After ruling over most of the philosophical views concerning introspective self-knowledge, Cartesian mentalism had shaped early experimental psychology. It is comprehensible, then, that philosophers, psychologists and neuroscientists were bewildered by phenomena that appeared to be *mental* but extended beyond the sphere of awareness and conscious control.

Two strategies were adopted to reconcile the existence of supposed unconscious mental phenomena with the consciousness-dependent conception of mind (see Livingstone Smith, 1999). The first option consisted in denying that such phenomena were genuinely unconscious; the evidence for unconscious mental states was reinterpreted as evidence for the possibility of a 'dissociation' or 'splitting' or 'doubling' of consciousness: 'the total possible consciousness may be split into parts which coexist but mutually ignore each other' (James, 1950, p. 206). The second option consisted in denying that such phenomena were genuinely mental; the evidence for the existence of unconscious mental states was reconceptualized as evidence for neurophysiological dispositions for genuinely (i.e., conscious) mental states.[9]

When, in the last decade of the nineteenth century, Freud intervened in the dispute on the unconscious, he took sides against the predominant 'consciousness-centric' mentalism in favor of the reality of occurrent and intrinsically unconscious mental events; and originally developed the concept of unconscious in two particular directions.

In the first place, Freud puts forward the idea of *sexuality* of the unconscious. At the heart of the unconscious is what Freud calls 'drive' (*Trieb*).[10]

[9] As Livingstone Smith (1999) notes, the two strategies are still options in current Anglo-American philosophy. John Searle has recast the dispositionalist approach to unconscious mental states (see below); whereas the so-called 'partitionist' approach to self-deception has revived the dissociationist option. See Davidson (1982), Pears (1982).

[10] Freud used both the German terms 'Trieb' and 'Instinkt'. However, as Schmidt-Hellerau (2005) has noted, although Strachey translated Freud's term 'Trieb' as 'instinct', it is more accurately translated as 'drive' and is to be distinguished from instinct, for which the German word is 'Instinkt'.

This is a relatively indeterminate pressure 'originating in a bodily source and aiming toward an object through which the drive is able to achieve its aim' (Freud, 1915, p. 122). Initially, Freud's drive theory comprised sexual drives and self-preservative or ego drives; but after Freud's (1923) metapsychological revisions, sex and aggression became the basic drives.

In a cultural-historical perspective, the idea of a sexuality of the unconscious is an important step in a materialist and pessimistic process of revision of the anthropological model of nineteenth-century middle-class ethics—a model that rested on the assumption of a full responsibility of the individuals toward an inner life consisting of conscious and self-transparent intentions. Such a revision was fostered, on the one hand, by Darwinian naturalism and the medical biologism of the nineteenth century; and on the other, by an anthropology of the crisis of Reason which, originating from Romanticism and the skeptical thought of previous centuries (above all Hume's), had found its main theorists in Schopenhauer and Nietzsche.

Freud's theory of the unconscious, therefore, offers a psychological formulation of themes that had previously been expressed mainly in philosophy and literature. But with a significant difference: he strove hard to contain the most disruptive aspects of the crisis of the traditional image of human rationality by proposing a version of it in which, though in the context of a non-optimistic conception of human nature, he suggested that neurotic suffering is connected to the mismanagement of the relationships with the unconscious, resulting in unhealthy forms of self-repression. In this perspective, psychoanalytic therapy offered the attractive perspective of a better management of the relationships between the unconscious and consciousness, encouraging in the conscious part of the ego the capacity to govern one's relationships with the unconscious in a more conscious and rational manner. To put it in the terms of the well-known formula with which Thomas Mann summarized Freud's thought: 'Wo *Es* war, soll *Ich* warden' (Where *id* was, *ego* shall be).

By contrast, in a scientific perspective, the conceptualization of sexuality in terms of drives is definitely the most timeworn part of Freud's work and, since the bioenergetic model of the mind is the main theoretical assumption of Freud's psychoanalysis, is not a minor shortcoming.

The debate on the concept of instinct with its variations (tropisms, reflexes, drives, etc.) runs throughout the history of psychology. Already

under attack since the 1920s, the idea of instinct as a definite quantity of energy that 'discharges itself' (according to Lorenz's famous drive-discharge or 'hydraulic' model of instinctual motivation) waned in the 1950s on both the biological front, by virtue of the British school of ethology's study of behavior in terms of signals (Griffiths, 2004a), and the experimental front, in relation first to the development of studies on the mechanisms of learning, and subsequently to the appearance on the scene of information theory (with cybernetics and systems theories, and later with computer science). Since the 1960s, with the rise of cognitivism, psychological functions (a concept that Freud did not possess) have been defined in terms of signals and information.[11]

However, as early as the 1930s and 1940s we find in the psychoanalytic field as well an implicit crisis of the centrality of drive in the theory of object relations, founded by Alice and Michael Balint, and later developed mainly by William Fairbairn, Donald Winnicott and John Bowlby. In Freud the newborn's original state is characterized as a condition of *primary narcissism*, that is, a sort of monadic self-sufficiency from which infants emerge only under the urge of their primitive sexual drives. The love attachment to the mother is, therefore, secondary to the '*Besetzung*' ('cathexis') of the mother's breast by the libidinal energy in its original oral modality. According to the theory of object relations, by contrast, the 'object-seeking' (i.e., the quest for the relationship with the caregiver) is not secondary to the need of drive discharge; it is *primary*, and the role of drives is therefore drastically downsized (Balint, 1965). The criticism of the concept of drive will become explicit with Bowlby's theory of attachment. Finally, the most systematic and radical attack against Freud's idea of instinct is launched in the USA, in the framework of the influence of David Rapaport's school. Since the 1980s, the idea that Freud's theory of instinctual drives can no longer be defended in light of scientific findings has become a recurring theme in the psychoanalytic debate (Holt, 1989; Macmillan, 1997).

The second main feature of the Freudian concept of the unconscious is that 'unbearable' mental contents are unconscious in that they are *repressed*, that is, actively excluded from consciousness owing to the

[11] Kurt Lewin was the first to introduce the concept of psychological function between 1930s and 1940s, and under Ernst Cassirer's direct influence.

unconscious activation of defensive mechanisms. This concept, too, is timeworn. For it has now become clear that the phenomenon that Freud called '*Verdrängung*' ('repression')—that is, the total (and irreversible, unless appeal is made to specific techniques such as hypnosis or psychoanalytic treatment) erasure of memories of traumatic experiences from our conscious minds—even if it exists, is extremely rare. In addition, there is no experimental evidence that such a phenomenon is in itself sufficient to produce long-lasting negative effects on an individual's mental stability (Loftus & Ketcham, 1994).

After Freud, however, a weaker sense of 'repression' established itself, even at the commonsense level. This is a meaning that we find in quite usual sentences such as, for example, 'I only remembered the date when it was already too late'. If someone said, as is now very common to say, that I 'repressed' the recollection of that date, what she would mean is not that I erased the date from memory but rather that I temporarily set it aside (in an *interested* manner: I did not want to remember). This weaker sense of repression is highly significant to the extent that it is consistent with a view of consciousness that is different from Freud's. On Freud's view, on one side there is consciousness (well separated from the unconscious), on the other, the 'stumbles' of consciousness (caused by the unconscious's infiltrating into the consciousness). Such stumbles occur only in a few exceptional or anomalous cases such as, precisely, repressions (in the strong sense).[12] But today we realize, thus deepening and confirming Freud's idea but also making it more radical, that our consciousness is *globally* permeated by the unconscious, namely, by a multitude of defensive strategies that are closely akin to repressions (in the weak sense).[13] The line separating the conscious and the unconscious thus becomes blurred and uncertain. In other words, distractions and selective records of events, memory lapses, temporary 'repressions', perceptual and conceptual scotomas, incomplete awareness, the dismissal of pieces of knowledge, rationalizations, amnesias and the partial or radical alteration of memories (including the invention of memories) turn out to be the very tapestry of our mental life.

[12] Manson rightly notes that in Freudian psychoanalysis the hypothesis that consciousness is not a necessary condition of mentality is applied only to 'a few exceptional or anomalous cases (slips, neuroses, etc.), and relative to a conception of mind as paradigmatically conscious' (Manson, 2000, p. 163).

[13] This can be clearly seen in Bowlby's (1980) theory of selective exclusion of information.

It can be maintained, therefore, that if in one respect the Freudian concept of repression is a museum piece, in another, it includes a reference to a persisting question, that of *bad faith*: our everyday thought processes are permeated by 'a self-apologetic defensiveness', by 'a systematic tendency toward self-deception' (Jervis, 2007, p. 150). In other words, ordinary human operativeness, in its cognitive and rational aspects, tends to conceal an underlying level of motivations, where emotional and affectional factors, originating from the primary interpersonal bonds in infancy, influence or even direct the rationalizations characteristic of the calculating thought in the context of interpersonal relationships.

This critical theme—the tendency of the mind to forge self-serving illusions—is one of the most important legacies of Freud's theory of the unconscious. Against the Cartesian conception of introspective consciousness as transparent awareness of our own mental processes and contents, Freud suggested that it is a construction packed with self-deceptions.

To begin with, Freud describes a *primary* self-deception when he sets up a contrast between the composite, non-monadical character of the mind and its unitary phenomenology. In the 'feeling of our own ego' (*Ichgefühl*) the ego (*das Ich*) 'appears to us as something autonomous and unitary, marked off distinctly from everything else' (Freud, 1929–1930, p. 13). But this appearance is deceptive: as a matter of fact, the ego is heterogeneous, heteronomous and secondary. In fact, it is the organized part of the id, which is totally unconscious and unstructured pulsionality, with which the ego is continuous 'without any sharp delimitation' and 'for which it serves as a kind of façade' (ibid.). Consequently, the ego is both the partial structure of disparate psychological functions and the apparatus that has, inter alia, the function of presenting to consciousness the immediate but illusory certainty of the existence of a mind that is fully conscious of itself, integrated, unitary, rational and controllable.[14]

Freud's hypothesis of a systematic tendency toward self-deception within our everyday thought processes has found a rich source of evidence in the experimental literature on self-knowledge. In social and group

[14] As we will see in Chap. 3, a great deal of cognitive science research today offers robust evidence for the hypothesis that the neurocomputational architecture of our minds is composite and decentralized, not monadic; and its appearing to consciousness as unitary is—as Freud suggested—a primary self-deception.

psychology we find experimental designs that prevent the participants from having any access to the real motivations (i.e., the real causes) of their behavior during the experiment. Despite being unaware of such motivations, they fabricate, in perfect good faith, *causal narratives* that have little or nothing to do with the real motivating factors—a fabrication that can be described as *rationalization*, or also as a non-clinical form of *confabulation*. Here, as we will see in more detail in Chap. 4, the everyday mechanisms of self-deception turn out to be more pervasive, articulated, various, and deep than the Viennese thinker imagined. In this respect, the current psychology of the unconscious is much more Freudian than Freud.

* * *

If the theme of bad faith is the strength of Freud's concept of the unconscious, the relationship between consciousness and the unconscious, as it unfolds in Freud's theory of repression, is the clue to the main difference between the psychoanalytic unconscious and the unconscious states and processes posited by cognitive scientists.

Today Freud's view of the relations between conscious and unconscious mind is the ground of the conception of consciousness dominant in the folk culture concerning the mind; actually, it may be said that the latter is a largely psychoanalytic culture (Castel, 1973; Moscovici, 2007). And of course, this culture represents an advance on the Cartesian thesis of the transparency of the mind, which informed the image of human beings typical of the nineteenth-century middle-class ethics challenged by Freud. If the Victorian anthropology was dominated by the idea of consciousness (and conscious agency) such that a person could say 'If I did it, it is *evidently* because I chose it, because I wanted to do it', in the folk psychoanalytic culture of the mind it is realized that people are tossed about by instances which they do not always control very well, such that at times anyone can legitimately say 'I did it, but I hardly know why', thus implying that one is at least somewhat at the mercy of one's own psychological world.

Thus the folk psychoanalytic culture of the mind makes an important correction to the idea of a mind consisting in conscious and self-transparent intentions. But it is only a partial correction. In this culture the most evident limitation of Freud's view of the unconscious still holds:

his definition of the unconscious is still given by its *difference* from—and in some respects also *dependence* upon—the definition of consciousness; the latter is taken as a self-evident, primary datum, although it is then criticized and diminished in comparison with the traditional, idealistic view: 'What is meant by consciousness we need not to discuss; it is beyond all doubt' (Freud, 1933, p. 70).

It must be specified that if Freud preserves the primacy of consciousness, it is not because he develops a phenomenology where this consciousness is the methodological ground for the investigation of reality. In other words, Freud does not develop a theory of subjectivity at all, nor even a theory of knowledge that starts from subjectivity. The very concept of subjectivity, or 'experientiality', was not part of Freud's toolkit. His way of theorizing, more than neglecting the subjective dimension, tends to translate it into objective terms, like a collection of mechanisms and energies. Described by means of a highly original and sometimes informally imaginative idiom, the places, forces and events in the Freudian mind never cease to be markedly reified. All Freud's thought is characterized by the influence of positivism: the mind is a world of facts or even objects.

Freud then claims, in accordance with a positivistic objectivism, to describe neurobiological mechanisms as constitutive of the mind. However, although these mechanisms aim to explain many dimensions of affectional and emotional life, they are not supposed to explain consciousness. As just mentioned, Freudian adult (self-)consciousness is, in spite of the dynamic unconscious, once more 'assumed' or 'given'. So we find in Freud's thought the persistence of a partial endorsement of the Cartesian model of the subject, which postulates a perturbing corporeal influence on the mind (*les passions de l'âme*) but also rigidly safeguards a primary (and, in Descartes, transcendent) principle of human rational awareness.[15]

Following a similar methodological approach, Freud deals with the problem of the very young child's mind by its subtraction from the adult mind. Neither Freud nor his contemporaries ever investigated the infant's mind in its autonomous genesis, or according to its own (viz. 'bottom up') standards—possibly taking advantage of animal behavior research. Infants are always viewed and evaluated not from the standpoint of their

[15] In Chap. 5, we will return to Freud's partial endorsement of the Cartesian model of the subject.

world but from the standpoint of the adult (see Peterfreund, 1978). After all, the need for a bottom-up study of the infant's consciousness, taking a decentralized standpoint and making hypotheses that do not reflect the phenomenological categories of adult self-consciousness, has become established only since the 1920s, thanks to the work of Jean Piaget.

Consequently, and as already noticed in Sect. 2.1.3, the Freudian unconscious turns out to be essentially an enlargement, or extension, of a psychology—folk psychology—hinged on the idea of a person who is able to have conscious mental experiences. In Freud's second topography (the id-ego-superego model of the mind), as noted by Laplanche and Pontalis, the model is no longer one borrowed from the physical sciences, as it was in the case of first topographical conceptualization of the psychical apparatus (the conscious mind, the preconscious and the unconscious mind), but is instead shot through with anthropomorphism:

> …the intrasubjective field tends to be conceived of after the fashion of intersubjective relations, and the systems are pictured as relatively autonomous persons-within-the-person (the super-ego, for instance, is said to behave in a sadistic way toward the ego). To this extent then, the scientific theory of the psychical apparatus tends to resemble the way the subject comprehends and perhaps even constructs himself in his phantasy-life. (Laplanche & Pontalis, 1973, p. 452)

And yet, in pointing out Freud's difficulty in emancipating the sphere of the mental from consciousness, it is not to be omitted that in his later years his effort to move beyond the consciousness-centric mentalistic framework became more radical. This can be seen from the fact that Freud came to think that the id reigns over our whole mental life. As a consequence, consciousness lost its importance, as did the organized part of the mind, that is, the ego. So much so that he dramatically maintains that the id uses the ego as a kind of façade.

* * *

According to a number of philosophers and psychoanalysts, Freud's preservation of the folk-psychological conception of mind is not a flaw of his theory. In this perspective, the grounds for psychoanalysis 'lie in its offering

a unified explanation for phenomena (dreaming, psychopathology, mental conflict, sexuality, and so on) that commonsense psychology is unable, or poorly equipped, to explain' (Gardner, 1999, p. 684). This perspective in the analytic philosophy of psychoanalysis originates with Davidson (1982, 1985). In his view, the personal level is autonomous and different from the subpersonal one, and is to be studied by means of different methods—*hermeneutics* takes the place of the quest for natural laws. This approach is the basis of a defense of psychoanalysis against well-known methodological objections (e.g., Grünbaum, 1984). Like folk-psychological explanations, psychoanalytic explanations need not meet the epistemological and methodological requirements of experimental science (e.g., Hopkins, 1988; Wollheim, 1993).

This attempt to abandon Freud's positivistic naturalism and reconstruct psychoanalysis on hermeneutic grounds has a very long story. In the 1970s, an influential version of this project was initiated by a number of psychoanalysts of Rapaport's school: especially George Klein and, close to his ideas, Merton Gill and Roy Schafer. According to these psychoanalysts, the 'biologistic' Freud is no longer defensible, and the whole Freudian metapsychology is to be declared obsolete, owing to its association to the drive discharge theory. By contrast, the psychoanalytic clinical theory must be reevaluated insofar as it rests on the intentionality of the interpretive process (see Gill & Holzman, 1976).

This 'clinical theory versus metapsychology' argument, however, tries to regenerate psychoanalysis by renouncing its main legacy. Although Freud's drive theory can no longer be defended in light of scientific findings, it is to be emphasized that it is precisely what ensures, for psychoanalytic theory, the strength of its criticism of the traditional idealistic illusions about the claim of self-legitimation made by rational consciousness. In this perspective, the metaphor of drive refers to something real, that is, the force of the 'matter' inside our mind. The Freudian hypothesis of a biological component that is constitutive of mental life, to the extent that it conceives human awareness as continuously 'tricked' and 'caught unprepared' by its biological dimension, rules out the possibility that mental life can regain its center in the free intentionality of consciousness. On the contrary, a psychoanalytic hermeneutics aimed entirely at an insistence on the theme of meaning, that is, on the intentional directing of consciousness (e.g., in giving sense to the object of interpretation), at

the expense of the biological theme of drive dynamics, runs the risk of surreptitiously reintroducing the traditional and pre-Freudian picture of the conscious subject as primary subject. For the subjectivity theorized by the hermeneuticists is 'inevitably *intentionalizing* rather than *intentionalized* by the cunning of the unconscious and the biological backdrop of the mind' (Jervis, 1989, p. 164).[16]

Furthermore, and closely related to the point just made, the psychoanalytic hermeneuticism tends—especially if under the influence of poststructuralist or deconstructionist ideas—to take the form of interpretive conventionalism. Interpretation is then ultimately committed to the freedom of deciding the meaning of the text on the strength of the agreement reached by the participants to the interpretive operation. But in this way the problems of truth and reality, of adequacy and confirmation, tend to disappear, being replaced by a freely creative narrativism of postmodern type (Eagle, 2003; Goldberg, 1984). This dismantles the project of *demystification*—the systematic search for self-deception and the uncovering of underlying truth—which is at the core of the critical tradition to which Freud belongs. Such an 'unmasking trend' has been part of European thought from La Rochefoucauld through Enlightenment philosophers, Marx, Nietzsche, and Ibsen (Ellenberger, 1970, p. 537).[17]

Finally, 40 years after George Klein's *Psychoanalytic Theory* (1976), it must be admitted that the 'clinical theory versus metapsychology' project has not produced results capable of breathing new life into psychoanalysis as treatment. In other words, now it is not only metapsychology that is in crisis, as in the 1970s, but clinical theory as well. All over the world, psychoanalytic treatment, even in its less doctrinaire and more well-structured and intelligent developments (and hence also bearing Rapaport's school in mind), has lost credibility and the market. So at this point, we are moving within a context of ideas that is not only post-Freudian but also post-psychoanalytic (Jervis, 2002, p. 49).

* * *

[16] All translations from Italian texts—unless otherwise indicated—are ours.

[17] See also Ricoeur (1970), who portrays Freud as a *philosophe du soupçon*, whose name should be associated to those of Marx and Nietzsche.

Our rejection of the autonomist project of putting psychoanalysis on exclusively hermeneutic basis paves the way for an exploration of the possibility that the psychoanalytic critique of the subject can be more rigorously conceptualized by replacing Freud's positivistic naturalism with a cognitive-evolutionary form of naturalism. This requires us to go beyond psychoanalysis in order to move into the field of dynamic psychology, an academic discipline that aims to analyze and develop psychoanalytic theories in close contact with cognitive sciences, especially all the systematic investigations that went to great pains to shed light on the ways in which the biological is constitutive of mental life.

Dynamic psychology, then, picks up the critical content of Freud's psychoanalysis: it is being aimed at the demystification of the spontaneous illusions about the existence of a subject that is *primarily* unitary, coherent, compact, self-justified and somehow 'noble'. But now, the unconscious turns into the *subpersonal* level of analysis of cognitive science. As we will see, in attachment theory psychoanalysis redefines itself within an ethological and evolutionary framework and posits 'a *cognitive* unconscious of beliefs, self, object and interactional *representations*, and implicit assumptions and expectations regarding how significant others will behave toward oneself' (Eagle, 2011, p. 130; emphasis added). And in the information-processing frame of reference, consciousness is no longer an unquestionable assumption, a non-negotiable given fact; the concept of the cognitive unconscious is no longer patterned, as in Freud, after the concept of conscious mind. Rather, cognitive science's subpersonal processes show features different from those of consciousness: whereas the latter seems to be unitary, serial, language-like, and receptive to global properties, the former are multiple, parallel, non-linguistic, and oriented to the processing of local properties.

It might be objected that such a claim needs to be 'calibrated' bearing in mind that in some cases the cognitive-science unconscious processes, too, are a bit too akin to the idea that is intuitive to the folk. Some cognitive-science models and specifically Fodor's computational-representational theory of mind tend to reproduce the operation of conscious thought processes. Thus, for example, Fodor's theory assumes that there are symbols with content, or that there is a computational state in correspondence to each folk state of belief, or even that cognitive processes (including the

perceptual ones) can be assimilated to deductive chains. On the whole, however, it can be affirmed that, given the very way of conceiving the mind in cognitive science as something halfway between the personal sphere of the first-person phenomenology and the non-personal domain of neurobiological events, the cognitive unconscious does not faithfully reflect the conscious level, and that the models of the unconscious adhering more closely to the structure of awareness are likely to belong more to the past of cognitive sciences than to their present.

2.3 The Unconscious in Cognitive Science: A Critical Discussion

We concluded the previous section highlighting the advantage of a dynamic psychology driven by cognitive sciences against the hermeneutical approach to psychoanalysis. In this perspective, the concept of the unconscious is no longer patterned after the idea of a person who is able to have conscious mental experiences.

This claim, however, might be challenged by the convergence of the above discussed interface problem and the contention that the cognitive unconscious (the subpersonal realm) is actually not mental. Since the assessment of the latter issue depends on what the application criteria for the concept of mental are, the issue has been called the 'mark of the mental' problem (see Armstrong, 1968). As we pointed out in Sect. 2.1.3, these two problems are linked, since if the cognitive unconscious is not mental after all, then interfacing computational psychology with ordinary psychological explanation turns out to be very hard (it is not easier than interfacing neuroscience and common sense explanation). Therefore, we must at least sketch a solution of both the interface and the mark of the mental problems. This is the goal of this section.

The subpersonal processes and representations postulated in cognitive science are taken to be mental entities. Yet this assumption is far from trivial. Indeed, one might raise the following objection: even though the computational account of a cognitive process is *not* a neurobiological account, what makes it *mental*? Would it not be more appropriate to say that cognitive science investigates, at a functional-computational level,

those *cerebral* processes that make our mental capacities possible? In other words, one could argue that, in cognitive science, only the *explanandum* is mental, not the *explanans*. Edge detection, for instance, is an operation entirely performed by certain neural circuits in the early vision area. Therefore, the computational theory of vision explains a mental faculty in terms of certain brain processes.

On this view, cognitive models provide high-level descriptions of neurophysiological processes, although it seems sensible to say that something deserves the predicate 'mental' only if the *person* is involved. When, for instance, one tries to explain language understanding in terms of the construction of propositional representations (or, for that matter, of sensorimotor simulations), the person disappears. The 'subject' of these processes is the subpersonal mind, but, from a metaphysical point of view, it is questionable whether the so-called 'subpersonal mind' is something different from the brain. On the one hand, we do not want to reify or objectify the mind, lest we fall into dualism; on the other hand, the loss of a linkage with conscious phenomena, or their marginalization, seems to result in a loss of the sense of what mind is *for us*. In the effort to understand the mind, the mind itself gets lost.

An influential version of this criticism has been formulated by the distinguished philosopher John Searle, who has questioned the assumption that subpersonal processes are mental. Let us see how.

2.3.1 Searle Against the Cognitive Unconscious

According to Searle (1990, 1992, Chap. 7), the ordinary concept of the unconscious mixes two categories that should be sharply distinguished. On the one hand, there are the (contents of) states that *cannot* become conscious; on the other, there is what can emerge into consciousness. The first kind of unconscious, which Searle suggests calling '*non-conscious*', is not mental at all; there is no reason to regard it as mental in any respect. Non-conscious mental processes, such as the processes typically posited in cognitive science, are operations of the brain, not of the mind. In this sense, according to Searle, there is no subpersonal mind. The mind is on an ontologically different plane from the brain processes that, as he says,

cause it. Therefore, Searle's thesis is that, properly speaking, something is unconscious (as opposed to merely non-conscious) if, while not being *currently* in consciousness, it has the possibility to become conscious; every mental state is either unconscious in this sense (potentially conscious) or actually conscious.

Searle's argument for this thesis is the so-called 'connection principle'. Two well-known notions are pivotal in order to grasp the argument: intrinsic intentionality and aspectual shape.

Following Brentano (1995), intentionality is the distinctive property of mental states: if something is a mental state, then it is an intentional state. Indeed, experiences, beliefs, desires, etcetera, are always experiences, beliefs and desires *of* something. Searle proposes two variations on Brentano's thesis: (1) not all mental states are intentional, since there are states, such as anxiety or nausea, which have no object; (2) *only* intentional mental states possess original or intrinsic intentionality, which is to say, if something possesses original/intrinsic intentionality, then it is a mental state: we speak as if intentionality were also a property of machines and artifacts, but in these cases *we* lend intentionality to systems which, *per se*, do not possess it at all. This is usually called 'as-if' intentionality. For instance, a biography of Charlie Parker can be said to be intentional, insofar as it is about Charlie Parker; clearly, however, it is the author of the biography who lent intentionality to the biography. The book has only as-if intentionality. From a slightly different point of view, this intentionality is merely *derivative*, insofar as it is inherited from the author.

As regard to aspectual shape, it can be defined as the perspective under which an object or state of things is given to a subject having a mental state. For instance, we can think of Aristotle as the greatest ancient philosopher, or as Alexander the Great's teacher: these are two intentional states, two thoughts, which are about one and the same object presented under a different aspectual shape. Since one cannot perceive or think of something without perceiving or thinking of it from one or another perspective–under certain aspects rather than others–, each mental state has an aspectual shape. For instance, when we are looking at a car, we see it from a certain point of observation, which allows us to see only certain aspects of it.

So far, so good. It is advantageous to present now Searle's argument in a systematic form:

(P1) Only intrinsic intentionality is mental: if something is a *bona fide* mental state, then is intrinsically intentional.[18]

(P2) Unconscious intentional states have intrinsic intentionality.

(P3) (Intrinsically) intentional states always have an aspectual shape, that is, they are about an object which is necessarily given under a certain perspective or mode of presentation.

P1 and P3 are regarded as unproblematic assumptions. P2 follows from P1, provided that unconscious intentional states are mental. Here Searle has in mind dispositional states such as most beliefs and desires. From P2 and P3 it follows that unconscious intentional states have an aspectual shape and from P1 and P3 it follows that mental states have an aspectual shape.

(P4) Aspectual shapes cannot be exhaustively described in third-person terms, for instance in a functional or behavioral vocabulary: 'There will always be an inferential gulf between the behavioral epistemic grounds for the presence of the aspect and the ontology of the aspect itself' (Searle, 1992, p. 158). In other words, the mode in which something is given to us cannot be completely predicted from behavioral, extrinsic facts.

(P5) The existence and character of an unconscious mental state are completely determined by neurological facts.

At this point, however, we get a contradiction. Indeed, the conjunction of premises P1–P4 entails that unconscious intentional states are *not* completely characterized in third-person terms, whereas P5 claims that an unconscious state is completely characterized by third-person facts, such as neurophysiological facts. Therefore, there is only one way to escape the contradiction, and this consists in claiming:

(C1) Necessarily, an unconscious intentional state *can* emerge to consciousness.

Indeed, the only way an unconscious mental state can have an aspectual shape consists in the possibility of being actual, that is, being conscious; by contrast, since neurophysiological states do not have any aspectual shape, they are non-intentional. According to Searle, the nature of unconscious states consists in possessing neurophysiological properties

[18] Since Searle does not think that *all* mental states are intentional, strictly speaking P1 should be expressed by saying 'If a mental state is intentional, then its intentionality is intrinsic'.

capable of causing subjective conscious thoughts or experiences. This conclusion has a corollary, which is exactly the thesis we are mainly concerned with here:

(C2) *Non-conscious states* (*states that cannot emerge into consciousness*) *are not mental*.

Indeed, if a state is non-conscious, then it has no first-person aspect (P4), therefore it has no aspectual shape (P3), and so is neither (intrinsically) intentional (P2), nor mental (P1).

The leading idea of the argument is that, for a mental state to have intrinsic intentionality, it must 'grasp' the world under this or that aspectual shape (under one or another perspective), but the possession of a perspective requires at least the possibility of being conscious. Therefore what is not even potentially conscious is not mental.

* * *

Many criticisms have been leveled against the connection principle. We mention three of them (Gennaro, 2012, pp. 22ff.):

1. The connection principle has a consequence that cannot be accepted: since many 'abnormal' psychological phenomena, due for instance to brain lesions or psychosis, cannot emerge into consciousness, they turn out to be non-mental (Rosenthal, 1990).
2. The connection principle predicts that several perceptual states, and in particular perceptual states devoted to motor control, are, *qua* non-conscious, non-intentional; by contrast, they are clearly intentional.
3. Searle's thesis draws a sharp divide between intentional phenomena and non-intentional phenomena. However, this clashes with the characteristic gradualism of natural phenomena. Since it is Searle himself who considers intentionality as a property with a biological basis, we should expect that intentionality, and subjectivity as well, emerge gradually (Shani, 2007).

All these objections aim to undermine, in different ways, the link established in premises P1–P3 among the properties of being mental, being intentional and possessing an aspectual shape. Notice that, while objections (2) and (3) involve an idea of intentionality quite different from the Searlean one

(see below), (1) seems to raise a difficulty internal to Searle's own point of view. Indeed Searle immediately addresses the problem raised by (1), which he dismisses by claiming that 'the possibility of interference by various forms of pathology does not alter the fact that any unconscious intentional state is the sort of thing that is in principle accessible to consciousness' (1992, p. 160). In other words, being, *qua* dispositional, potentially conscious is part and parcel of the nature of an unconscious mental state. However, it is not very clear how this reply is supposed to deal with the objection to the connection principle. Searle's idea, which seems to be that of a *de facto* impediment, such as a brain injury, does not change the metaphysical nature of a potentially conscious state. Therefore, it could be the case that, for instance, due to an injury, an unconscious state such as a belief can no longer (in nature) emerge into consciousness; yet, the metaphysical possibility of emerging into consciousness is not precluded. If this interpretation is correct, however, what is the *rationale* for denying that a given cerebral state can (metaphysically) emerge into consciousness? The distinction between non-conscious and unconscious states vanishes. Moreover, quite independently of this issue, what the objector wants to point out is that there are non-conscious states that can hardly be said to be non-mental. Searle's reply does not address *this* point.

Objection (1) and Searle's answer to it bring out a general difficulty: our intuitions about what is *mental* are far more insecure than we are inclined to believe; even though the notion of a cognitive unconscious faces the risk of dramatically broadening the concept of mind, Searle faces the opposite problem, as he ends up excluding from the mental domain something that should be presumably included. We return to the issue of the delimitation of the mental at the end of this section.

On the other hand, the existence of non-conscious intentional states evoked in objection (2), as well as the criticism of the characterization of intentionality as an all-or-nothing property put forward in objection (3), aim at questioning the thesis that intentionality is an exclusive property of paradigmatic mental states (propositional attitudes and perceptual experiences). Since intentionality is the property of being about an object or possessing content, nothing seems to prevent, at least *prima facie*, subpersonal states (computational rather than cerebral ones) from being bearers of such a property.[19] In fact, subpersonal computational

[19] Unless one is disposed to superimpose normative requirements on the notion of content. We shall not discuss this possibility here, because it seems to us not to fit well the Searlean view. Yet, if

states, *qua* representations, have an object or content; and it is customary to describe a neural firing pattern by saying that it carries information about such and such an object (or property) into the world.

Searle's strategy, in a nutshell, consists in arguing that the mental, necessarily, has a subjective component, by using intentionality as the link between mentality and subjectivity; indeed the crucial premise of the argument is P3, according to which each intrinsically intentional state has an aspectual shape, that is, is associated to a first-person perspective. However, this presupposes a restriction of the concept of intentionality that appears inconsistent with Searle's biological naturalism. The burden of the argument falls on the notion of *intrinsic*—as opposed to derivative—intentionality, but, despite its *prima facie* plausibility, the distinction between original (or intrinsic) and derivative (or as-if) intentionality cannot simply be taken for granted.

The rejection of the distinction between intrinsic intentionality and as-if intentionality is a major topic dealt with by Daniel Dennett, according to whom 'there is a continuum of cases of legitimate attributions [of intentionality], with no theoretically motivated threshold distinguishing the 'literal' from the 'metaphorical', or merely 'as-if', cases' (Dennett, 2009, p. 343). Dennett's theory is well known; we merely remind the reader of its crucial point: intentionality is an interpretive projection aiming to rationalize the behavior of a variety of systems (including ourselves), and there is no reason to think that *our* own supposed intentionality is special, since 'if it is not a miraculous or God-given property, it must have evolved over the aeons from ancestors with simpler cognitive equipment' (ibid. See also Dennett, 1987, Chap. 3 and passim).

One could reply that Dennett's position is controversial and cannot be directly used against the Searlean argument, to the extent that Dennett attacks the notion of intentionality as such. Indeed, he argues that *all* forms of intentionality are merely as-if. However, the notion of intrinsic intentionality fares not better, even if one has a (more) realist attitude toward intentionality. Taking intentionality seriously and considering it as a relation between a mental state and an object or state of things about which the

---

one wants to grant that intentionality is an *exclusive* property of folk mental states, the easiest way consists in taking an antinaturalist approach to mental phenomena (e.g., Voltolini, 2002).

mental state carries information does not force us to apply it only to folk mental states; in fact, it seems reasonable to claim that personal state intentionality somehow derives from subpersonal state intentionality, insofar as, for instance, my belief being about the cat Minou is explained, in the end, by the fact that there is a subpersonal (ultimately neurophysiological) state which codes Minou's being there. The idea is that the intentionality of at least some mental states basically depends on the active relation between an organism and its environment. This relation is realized by the construction of representations of objects and properties in the external world. This is, fundamentally, the brain's ability to code information about this or that object (or property) in the environment.

In short, Searle seems to face the following dilemma: either intentionality is nothing but a way of describing our practices of ascribing mental states, in which case all ascriptions are as-if attributions, and there is nothing physically or psychologically real in the idea of intentionality (this is Dennett's bent, at least in his more anti-realist moods); or by 'intentionality' we intend to refer to the property of carrying information about something (perhaps with some restrictions which are not easy to qualify properly), in which case intentionality is a property enjoyed by a variety of physical systems, including biological ones. Even if one opts for the second horn of the dilemma, the distinction between intrinsic and derivative intentionality does not work, with the unique exception of metaphorical uses, as in the earlier-mentioned case of the book being about, for example, Charlie Parker. As Robert van Gulick points out, cerebral processes underlying perceptual experience do not have beliefs about retinal images or eye movements, yet, 'from a functionalist perspective, they are nonetheless informed about them in a genuine and not merely as-if sense' (1995, p. 203).

2.3.2 Personal and Subpersonal in Dialectical Relationship

To recapitulate, Searle's argument against cognitive unconscious is not conclusive because the notion of intrinsic intentionality is not cogent. Since the argument (specifically P1) rests on such a notion, the whole

argument fails. On the other hand, even if we substitute intentionality for intrinsic intentionality, it is still false that only Searle's mental states have intentionality; some physical states can have it, too.

Yet, there is something in the connection principle and the related characterization of the unconscious that strikes us as somewhat plausible. It is an insight that can be expressed in the following way. The class of unconscious mental states must be reasonably restricted, otherwise any cerebral state or process even slightly characterizable in a computational/functional way, or playing a systematic causal role in the production of intelligent behavior, would belong to the category of the mental, including processes such as axon myelination, the releasing of neurotransmitters such as serotonin, activity of brainstem nuclei and the like (Damasio, 2010, p. 73). Indeed, all these processes play a causal role in our experience or in thinking but are not mental. Therefore, as we have already acknowledged (Sect. 2.1.2), some criteria of demarcation of the mental are required. Here are our suggestions.

According to a first proposal, something is mental if it plays a *direct explanatory role* in the theory of a given cognitive capacity or task (Paternoster, 2013). Take, for instance, edge detection carried out by off-center and on-center cells in the primary visual cortex. Since there is a computational description of this process playing a theoretical role in an explanatory model of a mental capacity (i.e., in Marr's computational theory of vision), then edge detection can be regarded as a mental process. By contrast, there is no cognitive theory in which, say, axon myelination or the releasing of serotonin plays an explanatory role; hence, they are not mental.

Admittedly, there are a couple of difficulties with this proposal. First, the notion of direct explanatory role is vague; second, mental character comes to depend on theories: if there is a change in the theoretical framework (such that the explanatory role changes, as well), a given process will see its ontological status change—from mental to non-mental or *vice-versa*. Arguably, this latter consequence could be acknowledged: after all, followers of Quine would not be shocked in the least by the claim that it is theory that determines ontology. Yet, this claim has a conventionalist, or even an anti-realist, flavor that will be rather unpalatable to many: it seems more reasonable to say that something is a theory of mind inasmuch as it is about the mind, not because it *constitutes* the mental.

Be that as it may, discounting the problem of vagueness (see below) the proposal based on the notion of direct explanatory role offers a way of reformulating the connection principle that can establish an explanatory bridge between personal and subpersonal.

The second proposal consists in *strengthening the link between the mental and the conscious* (Di Francesco & Piredda, 2012, Chap. 5). On this view, there are some forms of subpersonal information processing that, though not accessible to consciousness, can be considered as mental insofar as they have a specific relation with conscious processes; these subpersonal mental processes have a transparent and direct access to the personal mind.[20] With 'transparent' we mean that the subject is aware of the output of the subpersonal processes, but not of the processes themselves. We speak of *direct* access to avoid a slippery-slope overextension of the mind to neurobiological states that (for the reasons already noted by Searle) we do not want to call 'mental'. The notion of *transparency* here proposed is (freely) adapted from the extended mind debate, that is, the discussion of the thesis advanced by Andy Clark and David Chalmers (1998) according to which in certain cases the vehicle of cognition may be supplied by cultural and technological scaffoldings, such as external symbolic systems (words, numbers, maps, diagrams), technological props and the like (see Di Francesco, 2007; Di Francesco & Piredda, 2012, Chap. 5). In this context a process (even an 'extended' process) is taken to be transparent if it is invisible to the subject, who uses it in a fully unconscious and automatic way; yet, the results of the process must be accessible to the subject (even if the process itself is not).

Independently from the extended mind hypothesis, our idea is to take transparency and direct access to consciousness as a condition of mentality[21]: being transparent rather than being internal to the skull is what makes something mental. In this sense, transparency expresses the idea of a strong *integration* between the subjects' personal mind and their

[20] By 'personal mind' we mean the mind of the subject as is described by the personal level intentional/folk psychology—as composed of conscious, dispositional and Freudian unconscious states.

[21] In this case too, however, the difficulty of defining 'direct' access should be noted. Again, a possibility is to consider the overall structure of the theory of mental phenomena we are examining. If a subpersonal process is taken by the theory to produce a personal-mind content without further intermediary, then we can take its relation with personal mind as 'direct'.

other mental processes. Applied to the mark of the mental issue, this allows us to regain, for instance, those subpersonal states that are seemingly endowed with a representational content (e.g., Marr's 2½-D sketch, or perceptual processing in the ventral pathway) and, though being not directly accessible by the personal mind, are sufficiently integrated with personal processes. In sum, mentality requires integration between conscious and unconscious.

By putting these two proposals together, we are able to sketch a solution to the problem of the mark of the mental and, at the same time, suggest a way to address the interface problem. On the one hand, the notion of mental can be extended to incorporate subpersonal entities, provided that these are somewhat integrated with conscious processes (as is shown by the concept of transparency). On the other hand, in accordance with this criterion for the mental, we claim that the subpersonal approach has to be integrated with some references to the personal level, against the approaches that drop the link between personal and subpersonal. In the following section, we will see what kind of personal-level theories is apt for our project.

2.4 The Dynamic Unconscious in a Cognitive-Evolutionary Framework

We have seen that when we try to understand the relation between subpersonal and personal levels of psychological explanation, we face a dialectic between dependence and autonomy. If we consider the personal mind as completely autonomous, we fall into hermeneutics and into anti-naturalism, losing contact with the scientific developments. If we adopt a non-dialectical vision of the thesis of dependency, we end up adopting eliminative or reductive approaches that are at risk of losing the mental as their own object of study, replacing it with objects that belong to different levels of analysis. That being so, the wisest strategy may be to pursue a reflective equilibrium between dependence and autonomy, namely, working our way back and forth between the ordinary image of ourselves as conscious rational agents and the scientific conception

of ourselves as biochemically-implemented computational machines, by revising these two images wherever necessary so as to pursue the regulative ideal of a coherent self-conception.

A good example of a research area in which a dialectical relationship between personal and subpersonal levels of analysis turned out to be extremely fruitful is provided by the way in which a concept such as *attachment* allowed rethinking psychoanalysis in a cognitive-evolutionary framework. For this concept is very close to the personal level, taking shape in the context of a practical operative psychology rather than in systematic research. This is the above-mentioned attachment theory psychoanalysis, which is the psychodynamic framework within which our investigation on self-consciousness and identity will be carried out.

* * *

A distinguishing mark of the development of post-Freudian psychoanalysis is the focus on relational themes, especially on the forms of cognitive-affectional relationality of the very young child. The rise of attachment theory is part of this orientation. This theory hinges on two psychological constructs, motivation and attachment, which have served as bridges between dynamic psychology and cognitive sciences.

The central role that the concepts of motivation and attachment play in fostering an exchange between dynamic psychology and cognitive sciences must be viewed within the context of a deep revision of the anthropology underlying Freud's psychoanalysis. According to this traditional conception of human nature, whose paradigm can be found in Thomas Hobbes' political philosophy, individuality exists prior to relationality; sociality is a reality that comes 'after' individuality since it is a cultural product generated by the necessity to live together.

During the last decades, however, biology, sociology and behavioral economics have productively interacted with psychological sciences, making it increasingly clear that human sociality is not something that originates only from culture, but is rather a dimension that belongs to the definition of the human individual itself. According to this new anthropology, human sociality complies with certain natural predispositions; individuals are seen as bearers of a very complex suite of *motivations*, which are always, and have been from the beginning, *relational*. According

to Lichtenberg's (1989) well-known taxonomy, all these motivations give place to complex interactions between five 'motivational systems': the need to fulfill physiological requirements; the need for attachment and affiliation; the need for assertion and exploration; the need to react aversively through antagonism and/or withdrawal; the need for sensual and sexual pleasure. It is to be noticed, however, that the aversive-aggressive system is largely dependent on the assertive-explorative one, whereas the sensual-sexual system depends largely on the attachment-affiliation one. This led Jervis (2001) to suggest that the fundamental motivational systems may be only two: one dedicated to self-assertiveness and competition, and another aimed at prosociality and cooperation.

The Lichtenberg-Jervis model of motivational systems delivers an anthropology that is neither pessimist nor optimist: human beings are naturally inclined to competition, and sometimes destructivity, but also to forms of sociality, cooperation and even altruism (e.g., Bowles & Gintis, 2011). Freud saw the precarious situations of compromise between social repression and drive discharge as conflictual and sources of uneasiness. By contrast, the spontaneous situations of compromise that arise between the motivation to cooperate and the motivation to compete may turn out to be intelligent, well-organized and ingenious; and they are characterized not by uneasiness but by the generation of 'non-zero-sum' relationships.

The claim of the primary nature of sociality is then the anthropological foundation of the psychodynamics of object relations and attachment. The primordial psychological need of the very young child, around which his mental life gradually takes shape, is not—as Freud thought—the oral drive gratification, but rather the physical contact and the construction of protective and communicative interpersonal structures.

Attachment is the primary matrix of cooperation, and in Bowlby's (1969, 1973, 1980) theory, the dialectics between the attachment-affiliation system and the assertive-explorative system is the key to understanding the child's cognitive-affective development. For at the center of attachment theory is the relationship between the 'secure base' functions of the attachment figure and the individual's ability to explore the world and self, relatively free of anxiety. That is, one cannot comfortably engage in exploration (including self-exploration) without ties to other (i.e., a secure base).

* * *

The overcoming of the traditional philosophical and psychological view of the human individual as an isolated primary subject, *a priori* 'given' as autonomous, is the result of a contextualist and systemic perspective which puts the individual's psychological problems into the inter-individual and social context in which they arise and come to have sense. The theory of object relations seems to fully endorse this systemic approach to the study of relationality. As Donald Winnicott puts it, what makes sense is not considering the infant in itself, but the *mother-infant dyad*. But an epistemological caveat is in order here.

With the adoption of a systemic-relational perspective, psychology draws inspiration from trends currently dominant in biology and sociology. In biology the separation of the individual from the environment hardly makes sense. Both the developments of Darwin's theory and the modern concepts of equilibrium, adaptation, innate/acquired interrelation and ecological niche lead us to consider the individual-environment structure as a single systemic whole, where neither of the two poles is primary with respect to the other, and thus, also to consider the contrast between innate and acquired as obsolete. In animals as well as in human beings the development of the organism from the fertilized egg to reproduction and death consists in a series of structured interactions, each of which builds itself on the basis of the previous one, and each of which sees the interaction, on the one hand, of the onset of new 'environmental' signals, and on the other, the gradual opening of new 'inner' potentialities developed during the previous stages (Oyama, 2000a, 2000b; Oyama, Griffiths, & Gray, 2001).

In the case of the biological inspiration, the consideration of psychological phenomena in terms of equilibria, and hence of systemic interactions, has a naturalistic origin and is in continuity with William James' and John Dewey's functionalist school. Things change, however, when the systemic approach to the mind has a sociological origin. A forceful tendency has long existed in sociology and social psychology to attempt to make the investigation of human behavior more rigorously scientific by means of its *de-subjectivation*—hence the prevalent use of explanatory tools that have a structural-relational nature rather than dispositional-intentional one. We can already see this tendency at work in Talcott Parsons, with his turning Max Weber's typology of attitudes into a typology of role relations

(Gallino, 2006, p. 559). Such role relations are always structured, and in dynamic equilibrium, and can hence be considered in an implicitly systemic perspective. Similarly, the evolution of areas at the interface of psychology and sociology like, since the 1960s, symbolic interactionism and the work of Erving Goffman, resolutely points in the direction of a theory of the interactive construction of the description of the self and reality.

Now, what is primary in the systemic perspective is not the individual but the interaction, often viewed as a communicative dynamic field. It may happen, then, that sociological inspiration makes such an approach more radical and ends up dissolving the individual. The result is a form of sociologism that neglects the value of systemic naturalism, that is, wipes out any sense of an ecological perspective in which the human organism is *biologically* part of the environment before being sociologically and culturally part of it. This antinaturalistic sociologism gives place to a 'pure', disembodied relationalism, where the individual (the living and real information-processing organism) is reduced to a mere knot in the tangle of an organized field of influences or, more properly, messages.

A good example of this unwelcome outcome is provided by those forms of sociolinguistic constructivism which completely dismiss cognitive sciences, or seek to replace them with a 'psychology of the surface' which is relational and linguistic, such that there are no information-processing mechanisms, not even mental states and processes: these things are opaque and unproductive; only relations and language hold. On this view, psychological phenomena are produced in social interaction, and above all in the context of 'conversation', beyond which there is no mental process; mental processes are nothing but our conversational interactions. From here it is a short step to seeing persons not as the actors in or the agents of discourses, but rather as the products of the discursive practices themselves (e.g., Harré, 1986, 1987; and more recently, Carpendale & Lewis, 2006; Hutto, 2008). The self is, thus, entirely located within the public space of discourse.

Ironically, however, the suppression of the biological made by such antinaturalistic sociologism frustrates the very sense of integration that the systemic approach pursued, leading to the situation against which it aimed to struggle, that is, a conception in which (individual) biology and (social) relationality are split from each other, in that the former is deleted and the latter becomes all-encompassing.

Certainly, there is nothing in the theory of object relations that renders it ineluctably liable to such involution. Quite the contrary: it is wrong to think that if one speaks of the theory of object *relations*, then the theory is, as such, immediately *relational*. The idea of 'object relation' is not strictly and *in itself* an *interactionist* theory, let alone a *systemic* theory. The subject can still be seen as *primary* with respect to the object. In other words, we can still have a relation in the traditional sense, namely in a one-directional sense; the theory of object relations is not necessarily a relational theory in the strict sense, that is, a theory focused on the forms of an interactive dialectics that constantly generates new dynamic equilibria. That being the case, the different versions of the theory of object relations fit into different parts of the spectrum that from the classical conception of the subject seen as primary with respect to the object leads to the above rejected pure relationism. So we should not confuse and conflate the claims that minds are shaped by early interactions with others—and that much that goes on in our mind has to do with our relationships with others and representations of these relationships (all claims that we can find in the theory of attachment)—with the radical, social-constructivist claim that 'the basic unit of study' in psychoanalysis is not 'the individual as a separate entity' but 'an interactional field', which can be found in the relational theory of Stephen A. Mitchell (1988, p. 3).

* * *

In the psychodynamics of object relations and attachment, physical contact and the construction of protective and communicative interpersonal structures are regarded as the infant's primordial psychological needs, around which her mental life gradually takes form. This focus on the socio-communicative dimension is organically linked to an interest in the infant's subjectivity. The relational themes cannot be separated from the study of the ways in which the agents represent and experience their world environment.

To offer just one example, in attachment theory the attachment styles of children, correlated to the caregiving styles of parents, give place to 'internal working models', that is, the mental representations that originate from internalizing relational experiences with attachment figures. Internal working models of self and other in attachment relationships,

Bowlby claimed, help members of an attachment dyad (parent and child, or adult couple) to anticipate, interpret, and guide interactions with partners. This construct performed an important bridging function between attachment theory and cognitive psychology. In the third volume of *Attachment and loss* (Bowlby, 1980), for example, internal working models are defined in terms of Tulving's (1972) distinction between episodic and semantic memory. Children's conscious representations of what parents or others misleadingly told them may be stored as general propositions in the semantic memory system while the child's own (defensively excluded or segregated) memories of traumatic attachment experiences might be stored 'analogically' in the episodic memory system (see Bretherton & Munholland, 2008, p. 106).[22]

In the context of this interest for the infant's subjectivity, dynamic psychology appropriated a topic that is substantially absent in Freud, namely, the theme of the construction and defense of subjective identity.

The absence of the topic of identity from Freud's thought is tied to at least two reasons. First, as already mentioned, the concept of subjectivity is not part of Freud's toolkit; his way of theorizing more than neglecting the subjective dimension tends to translate it into objective terms, like a collection of mechanisms and energies. Second, in his vision of human itinerary of life, the demands of inner 'reclamation' prevail over the project orientations. In Freud's anthropological perspective, individuals can orient themselves—if they are capable of doing so—to acquiring a greater awareness of reality and healing from most of their neurotic symptomatology; however, they are not required to dispose themselves to a pathway of *self-realization*.[23]

In contrast, the problem of identity was at the core of William James' protophenomenology (see Wilshire, 1969). In the *Principles of Psychology* James famously distinguished three ways in which each of us, as an idiosyncratic *self*, grasps and defines our own identity: the physical, material aspects of the self (material self) associated to the bodily subjectivity; the subject's social identity (social self); and finally the spiritual self, namely, the individual's 'inner or subjective being, his psychic faculties or dispositions, taken concretely' (James, 1950,

[22] Section 4.3 will return to the connection between memory and inner working models.

[23] As is the case in Jung's *Psychological Types* (Jung, 1971).

p. 296). These psychological dispositions are identified by each of us in our subjective life. The spiritual self is grasped in a 'reflective process', and is the result of 'our abandoning the outward-looking point of view', to look inwardly instead (ibid.).[24]

However, if in the early history of scientific psychology the problem of identity was center stage, it almost disappears from psychological research thereafter. During the first half of the twentieth century, experimental psychology was almost exclusively concerned with basic problems concerning the structures of behavior and perception, and not with complex matters such as self-conscious subjectivity and identity. Animal psychology, for its part, although investigating issues concerning sociality, ranks and hierarchies, could not approach such a typically human issue as identity.

As a result, after James, matters of subjectivity and identity will not be addressed for a long time—namely, until the revitalization of consciousness research in cognitive science in the 1980s—almost exclusively in philosophy and sociology.[25] In sociology, this sometimes involved crossing the line to psychology, as in the case of Erving Goffman, who, by means of the construct of 'self-presentation' (Goffman, 1959), made it clear that each of us, without being aware of it, devotes a considerable amount of our energy to obtaining from others the continuous confirmation of the validity of our identity.[26] Also due to sociologists and psycho-sociologists—like George H. Mead—is the most important hypothesis about the nature of the feeling of subjective identity, namely, the idea that we see ourselves, and define ourselves, essentially through a creative process of internalizing the ways in which others see and define us (see Sect. 4.2.2).

For its part, dynamic psychology, too, was late in approaching this subject. It is necessary to mention here first Eric Erikson, although he was concerned with the acquisition of identity in childhood but not in

[24] In the next chapters the reference to James' classic reflection on the self, together to John Locke's, will turn out to be a compass that is still essential to navigating the debate on self-consciousness and identity.

[25] In philosophy matters of subjectivity have been pursued in both the analytical tradition—in particular in Wittgenstein's and Ryle's investigations on the logical grammar through which we conceptualize conscious experience—and the phenomenological tradition.

[26] Although Goffman has been read as an anti-realist about the self, in fact he views the self as 'a psychobiological process shaped by signs and symbols' (Schwalbe, 1993, p. 333). On the importation of the study of self-presentation into psychology, see Schlenker (2012).

infancy (see Sect. 5.2); and second Harry Stack Sullivan, who was the first to grasp the significance of the concept of self as developed by Mead, exploiting it in a psychological and more specifically psychodynamic context, that is, in the study of the interpersonal relationships and the inner dialectics associated with these relationships (see Sullivan, 1953).

It can be definitely said that the topic of identity plays a pivotal role in current psychological sciences and, with regard to three factors (see Jervis, 2006), this will be the subject of the next chapters.

The first factor concerns the theoretical psychology, and consists in the inextricable link between identity self-description and self-consciousness.

The second factor pertains to dynamic psychology and developmental psychology and consists in the fact that the construction of affectional life, in the course of infancy and, subsequently, throughout one's entire life, is closely linked to the construction of an identity that is well-defined and accepted as valid. The construction of a valid personal identity is inextricably linked to the construction and preservation of self-esteem; in turn, the issue of self-esteem cannot be separated from the issue of the 'solidity of the ego' (in Freud's sense), or the issue of the 'cohesion of the self' (in Heinz Kohut's sense).

The third factor concerns social psychology and consists of the fact that—as already mentioned—each of us constantly negotiates the validity of our identity in exchanges with other people.

3

Making the Self, I: Bodily Self-Consciousness

In this chapter, we begin to outline our own account of the self. We start with a criticism of the 'exclusion thesis', the claim that there is no room for something like the self in the natural order—a thesis that in modern philosophy goes back to Hume's and Kant's criticism of the Cartesian self. After having discharged the Humean eliminative approaches to the self, we turn to a critical examination of two different approaches to the theme of self-consciousness. The first perspective is that of analytic Kantianism, a line of thought that runs from Peter Strawson to Quassim Cassam; a tradition that has come into contact with cognitive sciences in the works of José Bermúdez. The second perspective is the project to provide a naturalistic version of the phenomenological claim that conscious experience entails self-consciousness, which has been pursued especially by Dan Zahavi.

Only certain specific aspects of these two perspectives will be taken into consideration and discussed in this chapter. First, the criticism of the Kantian thesis of the formal nature of the self will allow us to make the most of Bermúdez's emphasis on the nexus between self-consciousness and bodily awareness, identifying the ground of self-consciousness with certain bodily structures. However, we part

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_3

company with this philosopher when he considers such structures as instances of 'pre-reflective self-consciousness', intermingling his view with the neo-phenomenological approach to self-consciousness. In this latter perspective, pre-reflective self-consciousness is a *minimal* form of self-consciousness, which goes together with—and in a sense grounds—every conscious act. This is a kind of mental state that can already be found in the earliest stages of child development and is the basis of more cognitively advanced forms of self-consciousness. But we will argue that, so construed, the notion of pre-reflective self-consciousness is an empirically void construct, the artifact of a top-down approach to self-consciousness in which the philosopher's self-experience is idealistically taken as explanatory, instead of the phenomenon to be explained. Against this regressive tendency, our approach will be built around a clear-cut distinction between object-consciousness and self-consciousness. This allows bodily and psychological forms of self-consciousness to be seen as the result of a process of self-objectivation which requires conscious (but not self-conscious) representational activity.

In this framework, a case will be made for the hypothesis that the most minimal form of self-consciousness is *bodily* self-consciousness, the capacity to construct an analogical and imagistic representation of one's own body as an entire object, simultaneously taking this representation as a subject, that is, as an active source of the representation of itself.

As will be argued in Chap. 4, consciousness of the body as one's own body is necessary in order to construct self-consciousness as psychological self-awareness, and then narrative identity. Psychological self-description hinges on physical self-description, evolving from it through an interplay of mentalizing capacities, autobiographical memory and socio-communicative skills modulated by cultural variables. But our claim that the narrative self is neurocognitively and socially constructed does not prevent a defense of a *robust* view of it. In *eliminativist* versions of narrativism, made popular mainly by Daniel Dennett, the self simply does not exist as a causal efficacious entity: there is nothing but a confabulatory narrative elaborated by our brains to make sense of the chaotic flow of experience, and make social relations more effective. By contrast, we are proposing an approach to the narrative self that, in making an attempt

to mediate between neurocognitive, social-constructivist and narrativistic demands, radically dissents from the eliminativist standpoint.

The framework within which we pursue a viable story about how the narrative self is constructed after the onset of bodily self-awareness is a very definite interpretation of William James' well-known distinction between I and the Me. In Sect. 3.4 we argue that the I-self designates the very objectivation process that produces the Me-self—namely, it denotes the self-representing of the subject, where 'subject' refers to a psychobiological system.

3.1 The Disappearance of the Self

3.1.1 The Exclusion Thesis

The main topic of this section is the place of the self in the natural order. The question that we would like to address is the following: is there any room for the notion of self in the naturalistic picture of the world? This question stems from both a generalized suspicion about the theoretical utility of the very notion of the self and, above all, those interpretations of recent results in cognitive science that have led to eliminativism about the self, a position rooted in Hume's philosophy of mind and today pursued by philosophers such as Daniel C. Dennett and Thomas Metzinger.

We may label the negative answer to the question of the placement of the self in a naturalistic picture of the world as 'the exclusion thesis'. Recently, this thesis has gained supporters from all the provinces of philosophy. In his introduction to the *Oxford Handbook of the Self*, significantly entitled 'A diversity of selves', Shaun Gallagher (2011, p. 1) quotes three sources of the attribution of ontological inconsistency to the self. The first stems from the importance acquired by the embodied cognition approach to mental phenomena, which fosters the suspicion that classical conceptions of the self are still too Cartesian; the second makes reference to the poststructuralist deconstruction of traditional metaphysics; the third source is connected with a recent interest in Indian and Buddhist philosophy, with its explicit negation of the very existence of the self. Not surprisingly, the self/no-self debate appears

to be quite a chaotic battlefield, where embodied theories of the self stand shoulder-to-shoulder with social-narrativistic approaches, phenomenological theories of the minimal self, eliminative explanations of the illusion of the self, and so on and so forth.

To avoid the risk of getting lost in this philosophical maze, we shall introduce the exclusion thesis by taking advantage of (a contemporary reading of) two classical approaches on the subject: Hume's and Kant's. The exclusion thesis is an old issue, after all, which in modern philosophy originates from Hume's and Kant's criticism of the Cartesian ego. A criticism that makes clear what Quassim Cassam (1997) defines as the 'elusive' nature of the self.

Traditionally, one of the main sources of the exclusion thesis is the failure of what can be termed 'the Cartesian insight', the idea that each of us has introspective access to oneself as an inner mental entity. Both Hume and Kant criticized the Cartesian insight, and their criticism contributed to the idea of the puzzling nature of the self. In a nutshell, Descartes considered the self as a simple substance directly given in our conscious experience; Hume argued that we have no experience whatsoever of Descartes' simple ego, and Kant proposed taking it as a formal principle of identity—something that is presupposed by any mental experience, but cannot be directly experienced. In what follows, we shall briefly present Hume's and Kant's reasons for rejecting the Cartesian insight, trying to connect them with contemporary philosophical and scientific research, and in particular with Dennett's eliminativism from the Humean side and P.F. Strawson's conceptual analysis of Kant's views.

So, to start with Hume, it is well known that he denies that what we call 'our self' can ever be the object of direct awareness; the self can never be encountered in introspection: "…when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other […]. I never can catch *myself* at any time without a perception, and never can observe any thing but the perception" (Hume, 2000, p. 252). But what, then, is the mind, if one can have experience of it only as a place of disparate perceptions? Hume's answer is found in a famous passage:

> 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 propension 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. (Ibid., p. 253)

Hume is well aware that his conception of the illusory character of the unity of the mind owes us an explanation: 'What then gives us so great a propension to ascribe an identity to these successive perceptions, and to suppose ourselves possessed of an invariable and uninterrupted existence thro' the whole course of our lives?' (ibid.) Hume's own explanation of the genesis of the illusion of the (permanent) self is grounded in a mixture of clever philosophical analysis (mainly on the notion of identity) and old-fashioned associationist psychology, whose details we need not explore here. The result is the well-known claim that the self is just 'a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement' (ibid., p. 252). The unitary and continuous self is a fictional entity—perhaps a useful one, insofar as it gives our existence a sense of continuity, but metaphysically a fiction.[1] Note that Hume here is not denying *tout court* the existence of the self (which in fact is identified with a bundle of perceptions), but rather its substantial character. What Hume tries to do, then, is to explain the genesis of the illusion of the self's substantial nature, describing the psychological operations that transform a perpetual flux and movement into the appearance of a simple substance that persists as identical over time.

In fact, Hume was not happy with his own solution; in an *appendix* of the *Treatise of Human Nature* he lamented the incapacity of his theory to explain the unity of the mind—'to explain the principles, that unite our

[1] Thus, Freud follows Hume's lesson when he sets up a contrast between the composite, non-monadical character of the mind and its unitary phenomenology. See above, Sect. 2.2.

successive perceptions in our thought or consciousness' (ibid., p. 400).[2] This is a highly complex issue that cannot be tackled here, but this passage gives a first hint to the reasons for Hume's difficulties:

> The novelty and lack of uniformity that we find in our inner life make it difficult to see how Hume's appeal to resemblance and causality could possibly be enough to explain why we come to have an idea of an individual mind or self that endures through time. The true story must be at least more complicated than he allows. (Stroud, 1977, p. 127)

In the *Bounds of Sense* Strawson offers a reading of the problem that revolves around the difficulty that Hume faced in applying his analysis of the identity of material external objects in terms of resemblance and causation to the self as object of inner experience. Different perceptions can be considered perceptions of one persisting (external) body because of the occurrence of certain relations among them. If we try to apply the same strategy to the self, Strawson notes, we face a 'fatal lack of analogy' between the two cases (Strawson, 1966, p. 170). In the case of the perception of a body, it is possible to imagine how the involved relations may distinguish between the perception that should be associated to the body and other perceptions that should not. When we face the problem of self-identity, however, it is hard to see how such a distinction can be made (all the perceptions being, so to speak, in the same boat). From this, Strawson draws the conclusion that Hume's theory should be supplemented by the acknowledgment 'of the role of empirically applicable criteria of subject-identity' (ibid.). In fact, there is some consensus that what Hume's theory neglects is an *empirical* self, whose persistence grounds the associative mechanism that characterizes the working of the mind (see Fogelin, 1985, p. 101).[3] From our point of view, this thesis is interesting because of Strawson's reference to the absence in Hume of any appreciation of the role 'played by bodily identity in the empirical concept of a subject of experience'—a question also raised by Kant (Strawson, 1966, p. 169).

[2] We are painfully aware that Hume's (and Kant's) scholarship is endless, and probably every single sentence written by these philosophers is open to contentious interpretation. In what follows we select certain interpretations and offer readings of them that are patently driven by the purpose of paving the way for a more fine-grained analysis of the self based on contemporary empirical research.

[3] For a more detailed analysis, see Fogelin (2009, Chap. 6).

In any case, admitting that the true story of the birth of (the illusion of) the substantial self should be more complicated than that offered by Hume, how complicated should it be? Both a philosophical reflection based on conceptual analysis and an empirically informed philosophy of mind suggest that it should be quite complicated indeed.

3.1.2 Selfless Minds?

In Sect. 3.1.3 we will return to Strawson's descriptive metaphysics.[4] Presently, however, we are interested in the work of those philosophers who have made use of cognitive sciences to bolster the Humean skepticism about the self.

Nowadays we can rely on a great deal of cognitive science research that offers robust evidence for both the claim that our mind's architecture is heterogeneous and decentralized, and the Humean (and Freudian) hypothesis that in presenting itself to consciousness such apparatus stages a complex self-deception. These two ideas get a sophisticated philosophical and cognitive formulation in Dennett's narrative theory of personal identity. This philosopher famously rejects the hypothesis that there is, in some area of the brain, a place where 'everything comes together' for presentation to the inner self (the 'Cartesian Theater'). To this 'myth' Dennett opposes the Multiple Drafts model of consciousness according to which, at any instant, in any part of the brain, a multitude of 'fixations of content' occur (Dennett, 1991; Dennett & Akins, 2008; Dennett & Kinsbourne, 1992). The conscious character of these contents cannot be explained by their occurring in a *special* spatial or functional place (i.e., the Cartesian Theater), nor by their having a special format. Rather, it depends on what Dennett (2005) calls 'fame in the brain' or 'cerebral celebrity'. Like fame, consciousness is not an intrinsic property of the cerebral processes, but is more similar to 'political clout', a kind of influence that determines the extent to which a content affects the future development of other contents distributed all over the brain.

[4] According to Strawson (1959), *descriptive* metaphysics aims to describe the most general features of our conceptual scheme; *revisionary* metaphysics, in contrast, attempts to revise our ordinary way of thinking and our ordinary conceptual scheme.

In this framework, where heterogeneous and decentralized 'fixations of content' compete with each other in a quasi-chaotic process, Dennett—like Hume—finds no place for a self. No fixation of content has any sort of special 'personal status' and neither is it the case that a self is the sum of the elements involved in the messy competition he posits. There is no self given primarily and directly. And from this premise, Dennett infers that talk about the self as an 'I' directing the activity of the body is just fictional talk after all (a conclusion shared by Thomas Metzinger).[5]

On Dennett's eliminative view, a neuroscientific theory of consciousness must be a theory of how the *illusion* of the subject of consciousness arises (Dennett, 2005, p. 157). An amazing property of *Homo sapiens* is precisely the capacity to create a self: 'out of its brain it spins a web of words and deeds' (Dennett, 1991, p. 416). By means of this activity, the biological organism produces a narrative, and posits a 'center of narrative gravity' (Dennett, 1992). The narrative is the result of the working of a 'Joycean Machine':

> In our brains there is a cobbled-together collection of specialist brain circuits, which, thanks to a family of habits inculcated partly by culture and partly by individual self-exploration, conspire together to produce a more or less orderly, more or less effective, more or less well-designed virtual machine. (Dennett, 1991, p. 228)

The Joycean Machine is a software in the brain which creates the self, a 'virtual captain', a character described in internal and external discourse as the owner of the organism's mental states and as the actor of its actions and decisions, but who is in fact just a representational entity, not the real player in the game of human behavior. This inner character is just an abstraction, 'not a thing in the brain'. This seems to imply that a description of human agency that invokes the self cannot be an ultimate truth. The real explanation, which involves real causes, will be found at the level of the brain.

[5] Metzinger in his book *Being No One* writes: '…no such things as selves exist in the world. Nobody ever was or had a self' (2003, p. 1). Here again the idea is that there is no explanatory role played by the notion of self. For Metzinger even to speak of the self as an illusion may be too much: '…there is no one whose illusion the conscious self could be, no one who is confusing herself with anything' (p. 634). In this sense, he appears even more radical than Dennett.

Although Dennett's theory was developed in the early 1990s, and was necessarily partly metaphorical in character because of the relatively poor status of the empirical study of consciousness, more recent empirical research is consonant with it. A neurocomputational architecture largely compatible with Dennett's Multiple Drafts Model is the global neuronal workspace model of conscious access. There is now extensive evidence supporting such a model (e.g., Dehaene, 2014). Moreover, analyses of functional connectivity patterns in the human brain have demonstrated just the sort of neural architecture necessary to realize the main elements of a global broadcasting account. More specifically, these studies show the existence of two main neurocomputational spaces within the brain, each characterized by a distinct pattern of connectivity (see Dehaene & Changeux, 2011).

The first space is a processing network, composed of a set of parallel, distributed, and functionally specialized processors or modular subsystems subsumed by topologically distinct cortical domains with highly specific local or medium-range connections that encapsulate information relevant to its function. The subsystems compete with each other to access the second space, a neuronal global workspace consisting of a distributed set of cortical neurons with long-distance connections, particularly dense in the prefrontal, cingulate, and parietal regions, and which are capable of interconnecting the multiple specialized processors and can broadcast signals at the brain scale in a spontaneous and sudden manner. This global neuronal workspace breaks the modularity of the nervous system. When one of the subsystems accesses the global neuronal workspace, its outputs (sensory information including perceptions of the world, the deliverances of somatosensory systems, imagery, inner-speech and so on) are broadcast to an array of specialized executive, conceptual and affective consumer systems (e.g., systems that 'consume' the perceptual input to form judgments or make decisions). This broadcasting creates a global availability that is experienced as consciousness and results in reportability.

At least three features of a global broadcasting architecture are significant for Dennett (see Schneider, 2007). First, it assumes that the neurocognitive architecture underlying consciousness is a distributed computational system with no central controller. Second, it makes massive use of recursive functional decomposition, an indispensable requirement to get rid of any homunculus who, nestled in a sort of pineal gland, scans

the stream of consciousness. Third, it allows Dennett to hypothesize that the aforementioned political clout is achieved by 'reverberation' in a 'sustained amplification loop' of the winning contents (Dennett, 2005, pp. 135–136).

* * *

Dennett's view of the psychological self as a center of narrative gravity is a radically anti-realist narrative theory of the self. While brains (as well as human organisms endowed with a self) are 'things', the self itself is not; it is at most a useful fiction. As Schechtman (2011, p. 397) puts it, '[h]uman brains are narrative-generating machines and selves are the protagonists of the narratives they generate'—but these protagonists are no more real than literary figures such as Ishmael and Sherlock Holmes. As in the case of 'abstractions' such as the Equator or centers of gravity, it is useful (and sometimes indispensable) to interpret human behavior *as if* it were governed by an inner integrated system of decision and representation. But what in fact we face is a distributed society of subpersonal cognitive agencies (see Dennett, 1991, p. 367).

In the next pages, however, we will see that there are reasons to doubt that, strictly speaking, Dennett's anti-realist conclusion about the self is *necessitated* by cognitive sciences. For the moment, we limit ourselves to noticing that Dennett's anti-realism about the self is in fact based on *two* moves. The first consists in characterizing the self as a substantial inner entity that exerts a top-down control over behavior. The second is to claim that there is no scientific evidence whatsoever of the existence of entities of this kind. This claim may be supported by the further premise that science can explain the genesis of the illusion of the (existence of the) self. The second move is, in a clear sense, supported by science (at least adopting a standard view of what is meant by 'substantial inner entity'). But the first move is an independent (and dubious) philosophical claim.

The presence of both assumptions (and of the further premise) is not peculiar to Dennett, and seems to characterize many forms of eliminativism about the self. Thomas Metzinger's argument against the existence of selves is a good example. He starts by claiming that 'science offers conceptually clear models of functional mechanism which could parsimoniously explain the *integration* of individual property-representations into a

unified self-representation' (2011, p. 282). This means that, like Dennett, Metzinger has a theory designed to explain the genesis of the illusion of the self: (1) a set of psychobiological functions creates the representation of a substantial self, and (2) the subject is deceived into representing such a self as something real—failing to recognize its fictional character. This dynamical, bottom-up self-organization offers 'a new theoretical option' for the Humean philosopher (the 'bundle theorist'), an option that is reinforced by the fact that '[w]e just don't find a substantial self anywhere in the world, and nothing at the level of scientific facts determines our metaphysics in this way' (ibid., pp. 282–283).

It seems then that we face a strong case for anti-realism. All this presupposes, however, that if the self exists, it must be a persisting individual substance (see Tomasetta, 2015, pp. 138–139). In other words, if the only way to escape the Humean conclusions were to prove the existence of a 'substantial self', the absence of scientific evidence of the existence of such an entity, combined with a theory of the genesis of the illusion of its presence, would make the adoption of the Humean strategy our best option. What it is missing here is an argument to the effect that the only way to take selves as part of the furniture of the universe is to conceive them as substantial entities (as 'things' hidden in our head). The same criticism applies to Dennett's theory. The kind of self which is eradicated by Dennett's metaphor of fame in the brain is the neuronal counterpart of the Cartesian ego: a centralized brain system which acts as the Inner Boss at the top of the chain of command that flows from brain to behavior. What is implausible in light of the scientific findings invoked by Dennett is the existence of the Boss (the spectator of the Cartesian Theater), hidden in the brain, who is in charge and directs our behavior. In this connection, Dennett draws on the literature on self-organizing systems (of which the termite colony is a paradigm case) whose behavior looks organized and purposeful to the external eye, but is actually the emergent product of the joint operation of autonomous subcomponents. Nothing is said, however, against the possibility of conceiving the self as the product of bottom-up processes that engender 'self-governing' systems—that is, systems 'whose movements are orchestrated in part by a unified, self-centered, informational stream' as opposed to systems 'like an ant colony or a Brooksian robot in which behavior is the emergent product of the joint operation of a collection of non-intersecting informational streams' (Ismael, 2006, p. 352; see also Sect. 5.4 below).

It is important to note that our disagreement with the eliminativists is not merely verbal. We are not just arguing that the 'illusion' of the self may have causal powers in the sense in which a false belief may. This is obvious and irrelevant (even if it offers Dennett the surprising chance to deny that he never conceived the self as illusory—see Dennett, 2014). Indeed, one of the aims of this book is precisely to show that a great deal of solid empirical research underpins a 'robust' theory of the self that is compatible with realism, a theory that takes the self as a *causal* center of gravity, and not as a *façon de parler*.

We will return to this complex matter in Chaps. 5 and 6. For the moment we should take a step back and acknowledge the strength of Dennett's position. We should admit, in fact, that, although not necessitated by cognitive science findings, Dennett's anti-realism about the self *appears* at first sight to be the stance that is more congruent with them. Cognitive science begins with the idea of the fruitfulness of the bottom-up approach sketched in the Introduction. As the reader will recall, this approach proceeds bottom-up in the sense that attempts to reconstruct the evolution of the complex psychological functions underlying the self-conscious adult mind from more basic ones. It does not appeal to our introspective self-knowledge, but to all those disciplines that investigate the gradual construction of human self-awareness. The outcome is a criticism of the primacy of self-conscious subjectivity, which, far from being a primary, simple, given phenomenon, turns out to be the complex product of a process involving numerous neurocognitive and psychosocial components.

But at this point, we are faced with a dilemma. Either we endorse a radical anti-realism of the self and give up the personal-level image of a self-conscious agent, or we reject the eliminativist doubts about the self and outline a genuinely realist picture of the self. We opt for the second horn of the dilemma, and in this book we aim to show how it is possible to maintain a realist theory of the self, but, it should be noted, without sacrificing the merits of a bottom-up subpersonal strategy. In other words, we stand with Dennett in endorsing the bottom-up approach, but part company with him by defending a realist account of the self, as well as the utility and necessity of a picture of the mind which is sensitive to both the subpersonal and personal levels of analysis. As our previous reflections on the interface problem suggest, any strategy that introduces

a drastic divide between personal and subpersonal phenomena appears to be in contrast with current practice in important areas of the science of the mental.

3.1.3 Analytic Kantianism

In the previous section, we presented Hume's arguments against the possibility of finding within us the stable perception of a Cartesian immaterial substance, and the correlative vision of the self as a collection or bundle of perceptions. We also noted that the main difficulty of this position is the impossibility of explaining, within the Humean framework, what makes a series of experiences the experiences of a single specific person. We then critically examined Dennett's version of the Humean perspective based on an eliminative form of narrativism. Before starting to develop our positive view of self-consciousness and the self in Sect. 3.2, we still have to devote a few words to Kant's solution to Hume's problem.[6]

Kant's answer highlights the necessary link between experience and the subject to whom it belongs, and asserts that the awareness of the unity of one's own consciousness by the self appears as a prerequisite for the activities of categorization and experience of the world: 'Now no cognitions can occur in us, no connection and unity among them, without that unity of consciousness that precedes all data of the intuitions, and in relation to which all representation of objects is alone possible' (Kant, 1998, p. 232, A 107). In this sense there is no experience without a subject, no isolated perceptions/representations that only later come together in Humean bundles; from the beginning, they must be part of a single and integrated consciousness.

[6] Before starting our analysis a disclaimer is due. The Kantian themes that we are introducing (such as the notion of the 'I think', the necessary unity of apperception, the formal nature of the self) have been the object of endless scholarly disputes. Our ambition here, however, is not philological accuracy. Rather we are mainly interested in providing a link between the authoritative tradition of analytical Kantianism inaugurated by Peter Strawson (already mentioned in the previous pages) and contemporary reflections about the self. For a recent discussion of 'the project of advancing our understanding of the cognitive subject through examining Kant's theory of cognition' inaugurated by P. Strawson, see Kitcher (2011).

A first crucial element of Kant's theory is the advocacy of the thesis that the representations that I experience could not be *my* representations without what Kant calls the 'I think', which necessarily accompanies them—it must be possible that every state of consciousness is accompanied by the subject's 'apperception' of their belonging to the unity of his mind. In this perspective, the representations are given only in the unity of consciousness and are intrinsically and directly related to the subjectivity to which they belong.

Central to Kant's doctrine of the 'I think', then, is the thesis that, in order to attain the unity of the manifold of representation in one subject, a process of synthesis is necessary. By 'synthesis', Kant means 'the action of putting different representations together with each other and comprehending their manifoldness in one cognition'; such a process 'collects the elements for cognitions and unifies them into a certain content' (ibid., pp. 210–211, A78/B103). Synthesis starts with a multiplicity of representations and 'collects them with one another to produce a single further representation with cognitive content' (Pereboom, 2014, p. 5). In virtue of such synthesis, each representation can be accompanied by the 'I think', which is the *representation* of the synthetic unity of apperception that accompanies all our mental representation and guarantees the unity of the mind.

A second crucial element of Kant's theory is the *formal* nature assigned to the 'I think'. The 'I think' establishes a necessary condition that any form of subjectivity has to satisfy, a form of self-consciousness that accompanies every representation, but does not express the content of a single and concrete act of thought referring to the empirical subject. The 'I' of the 'I think' is, therefore, not an individual entity that we can perceive. To forget this leads to the error of rational psychology, that is, of confusing 'the unity of experience with the experience of unity' (Strawson, 1966, p. 162). The unity of consciousness does not lead to the experience of a unitary substance; the unitary subject of experience is not the object of his own experience. The 'I think' does not entail the perception of an object, but is the product of a function of the intellect. Kant's solution to Hume's problem is attained not by reference to the persistence of a thing or substance (this was in fact Descartes' error), but by appealing to the activity of the 'transcendental subject'. In the context of Kant's philosophy, then,

the 'I think' is not part of the empirical world; as Strawson puts it, the 'I think' of apperception is 'the tangential point of contact between the field of noumena and the world of appearance' (ibid., p. 173).

We need pursue Kant's theory of the synthetic unity of apperception no further, apart from noting that Kant's doctrine of the 'I think' rejects both Hume's 'causal' account of the unity of the self and the Lockean attempt to derive self-consciousness and the unity of the self from common contents in the flux of representations (ibid., p. 169; see also Rohlf, 2014, p. 29).[7] Kant agrees with Hume in criticizing Descartes' doctrine that solely on the basis of conscious experience we can know our existence as immaterial, simple and permanent thinking substances, capable of an existence independent of matter. Kant's explanation of the genesis of this illusion, however, is not based on a naturalistic psychological theory of the association of ideas, but rather on the dialectical error of mistaking, as we have seen, the necessary unity that accompanies all our experiences for the experience of a unitary substantial entity.

The formal nature of Kant's 'I think', which prevents the self from being part of the experienced natural world, is a point underlined and criticized by the phenomenological tradition. As Gallagher and Zahavi put it, Kant takes the self as 'a distinct principle of identity that stands apart from and above the stream of changing experiences' (2008, p. 200), and secures their unity and continuity. The self, as a persisting entity, should be distinguished by its changing properties and the Kantian 'formal' self does this job properly. But as a formal principle of identity, the self fails to be part of the world. The self is presupposed by any mental experience, but cannot be directly experienced—and this makes it an 'elusive' entity.

In the attempt to make the self a less elusive entity, we will focus our attention on Strawson's brilliant reformulation of Kant's criticism of the Cartesian self, which, albeit developed in terms of transcendental arguments, contains the seeds for some more empirically oriented lines of research that will be scrutinized in the next sections.

[7] As we shall see in Sect. 4.1.1, according to Locke, the concept of person is not an essence but rather a psychosocial attribute that is assigned to those subjects who possess a specific set of psychological capacities, which makes it possible the continuity of the self and the reflective appropriation of the subject's actions.

The great merit of Strawson's analysis of self-consciousness is to transform Kant's formal self into an inhabitant of our world. Strawson's key move is to argue for the idea that the criteria for the numerical identity of subjects of experience require some reference to the human body. In this sense, he takes a fundamental step toward rejecting the exclusion thesis, a step taken together with many phenomenologists, such as Husserl, Sartre and Merleau-Ponty, and which may find further support in contemporary science of the mind (see Cassam, 1997, p. 9).

Strawson addresses the issue of the elusiveness of the self both in *Individuals* and *The Bounds of Sense*. In the former book, he opens the chapter dedicated to the notion of *person* with the following question: 'Each of us distinguishes between himself on the one hand, and what is not himself or a state of himself on the other. What are the conditions of our making this distinction, and how are they fulfilled?' (Strawson, 1959, p. 87). Crucial elements in Strawson's answer include his criticisms of the 'Cartesian' and the 'no-ownership', or 'no-subject', theories of the self.

Indeed, a very important aspect of Strawson's analysis is his attempt to supply some missing steps in Kant's explanation of the genesis of the Cartesian insight. And the most important missing step (at least for our purposes) is the explicit appreciation of the fact that 'any use of the concept of a numerically identical subject of experiences persisting through time requires empirically applicable criteria of identity' (Strawson, 1966, p. 164). The Cartesian insight does not provide these kinds of criteria, and they cannot be found in the 'kind of connectedness of inner experiences provided for by the necessary unity of apperception' (ibid.). For Strawson, however, '[w]e *have* criteria of singularity and identity for subjects of experience (people, men)', so we should try to derive our criteria of individuation for souls or consciousnesses 'from the notion of singularity and identity of men and people' (something like: *one* person, *one* consciousness, *same* person, *same* consciousness). But this would be the 'suicide of rational psychology'. Drawing upon our criteria for individuating the empirical self is not something that Kant was prepared to do, however, and, according to Strawson, 'Kant's failure to press this point home is but an aspect of his neglect of the empirical concept of a subject of experience' (ibid., pp. 168–169).

The last remark leads us to the second connected point of Strawson's analysis that we would like to underline: 'the role played by bodily identity in the empirical concept of a subject of experience' (ibid.). According to Strawson, self-ascription of experience requires the existence of 'empirically applicable criteria of identity' for subjects of experience. A condition satisfied in actual practice 'by the fact that each of us is a corporeal object among corporeal objects, is indeed a man among men' (ibid., p. 102).

Strawson intended his reference to the corporeal dimension as an exercise of conceptual analysis, connected with the necessary conditions for the mastery of first personal pronouns: 'Our personal pronouns, the pronoun "I" included, have an empirical reference; and in some way such a reference must be secured if the general notion of ascribing experiences to a subject of them is to make sense' (ibid.).[8] His reading of the missing steps in Kant's explanation of the genesis of the Cartesian insight, however, may be developed in a naturalistic direction, such as that taken by a series of reflections stemming from scientific evidence of the strict connection between self-consciousness and bodily and environmental awareness developed by José Bermúdez. In fact, Bermúdez takes Kant as 'the philosopher who has had the clearest grip on the relation between self-awareness and awareness of the environment' (Bermúdez, 1998, p. 165), and discusses at length Strawson's reading of Kant in *The Bounds of Sense.* Strawson's Kant, according to Bermúdez, wants to ask the following question: 'What must hold for a series of thoughts and experiences to belong to a single, unitary self-conscious subject?' And Strawson's main claim is that 'no creature can count as a subject of experience unless it is capable of drawing certain very basic distinctions between its experiences and the objects of which they are experiences' (ibid., p. 166). What is required is the kind of consciousness that Strawson himself calls 'nonsolipsistic consciousness', a form of consciousness that allows the avoidance of the elusiveness of the self by connecting self-consciousness and bodily awareness.[9]

[8] In a similar vein John McDowell writes: 'We can say that the continuity of "consciousness" is intelligible only as a subjective take on something that has more to it than "consciousness" itself contains: on the career of an objective continuant, with which the subject of a continuous "consciousness" can identify itself' (1996, p. 101).

[9] 'I shall mean by a non-solipsistic consciousness, the consciousness of a being who has a use for the distinction between himself and his states on the one hand, and something not himself or a state of himself, of which he has experience, on the other' (Strawson, 1959, p. 69).

One main difference between Bermúdez's reading and Strawson's is that the latter, but not the former, subordinates nonsolipsistic consciousness to the possession of conceptual skills by parts of the subject. According to Bermúdez, we may speak of first-person thoughts (*I-thoughts*) endowed with non-conceptual (and *a fortiori* non-linguistic) content: 'Somatic proprioception and the structure of exteroceptive perceptual experience can be a source of nonconceptual first-person contents from the very beginning of life' (1998, p. 163). More precisely, Bermúdez, assuming that Strawson's transcendental argumentative style cannot really do without experimental evidence (see Bermúdez, 1995), envisages four domains of research within cognitive science that suggest the presence of this kind of non-conceptual content: (1) perceptual experience; (2) somatic proprioception (bodily self-awareness); (3) self-world dualism in spatial reasoning; (4) psychological interaction (Bermúdez, 2001, p. 134). The first kind of phenomenon refers to 'J. J. Gibson's great insights', according to which 'the very structure of visual perception contains *propriospecific* information about the self, as well as *exterospecific* information about the distal environment' (ibid., p. 135). The first-person perspective, then, is built into perceptual information—the self is experienced in perception as the boundary of the visual field: 'a moveable boundary that is responsive to the will' (ibid.). This kind of ecological self is the basis of self-awareness—together with what we may call 'the somatic self', the self of bodily self-awareness, or proprioception. Somatic proprioception is essential to giving us the sense of the self as a cause of action: it 'offers an awareness of the body as a spatially extended and bounded object that is responsive to the will', thereby contributing to the birth of the distinction between self and non-self (ibid.).

In Sect. 3.3 we will take a position on the onset of the self/non-self distinction that makes the most of Bermudez's emphasis on the nexus between self-consciousness and bodily awareness, but parts company with some crucial points of his line of reasoning.

3.2 The Bottom-Up Reconstruction of the Self

The criticism of eliminativism, the related need to maintain a sufficiently robust concept of self, and the focus on the bodily dimension of self-consciousness set the stage for our own account of the self. In this section, we begin to explore how it is possible to maintain a robust theory of the self within our naturalistic, bottom-up, systemic-relational framework.

In recent years, a bottom-up approach to the issue of self-consciousness, with its empirically informed account of the precursors of self-consciousness, has been much cultivated in theoretical psychology. Most approaches, however, assume a minimal form of self-consciousness as the very basis of cognitively more advanced forms of self-consciousness and construe this minimal self-consciousness as a 'pre-reflective self-consciousness', a tacit, non-intellectual sense of self that makes every conscious state a first-person phenomenal state (e.g., Gallagher & Zahavi, 2008, 2015; Prebble, Addis, & Tippett, 2013). As we will see in the next section, though, this is an empirically void construct, the artifact of a top-down approach to self-consciousness in which the philosopher's self-experience is (anti-naturalistically) taken as an *explanans* rather than an *explanandum*. Against this regressive shift, our approach is built around a clear-cut distinction between object-consciousness and self-consciousness.[10] This allows the possibility of viewing bodily and psychological forms of self-consciousness as the result of a process of self-objectivation, which requires conscious (but not self-conscious) activities of representation.

In this framework, the most minimal form of self-consciousness is a *bodily* self-consciousness, which consists in the capacity to construct an analogical and imagistic representation of one's own body as an entire object, simultaneously taking this representation as a subject, that is, as an active source of the representation of itself. This bodily self-consciousness—it will be argued—is needed as a foundation for the narrative identity.

[10] Self-consciousness could be regarded as a particular form of object consciousness, the consciousness of that particular object which is the self. Yet, several developmental stages are required to attain even the most elementary forms of self-consciousness.

Therefore we propose an account of narrative identity that parts company with those accounts that devote little attention to the role of the body in the narrative self-concept, or go to the extreme of stating that the narrative self is abstract and hence not embodied (for overviews of the debate on the relation between embodiment and narrative, see Brandon, 2014; Køster, 2016). On the other hand, we do not credit the hypothesis that the embodiment of the narrative self is provided by a pre-reflective self-consciousness viewed as a primitive, proprioceptive form of self-consciousness already in place from birth (as claimed by Zahavi, 2007, 2012, 2014).

* * *

The first, minimal condition required for the development of the self is the possession of a simple or primary *object consciousness*. Primary object consciousness is the mere experiencing of the objects and properties of the world, in virtue of the possession of representational capacities. Any organism endowed of perceptual and motor systems with a certain degree of complexity, that is, whose behavior is mediated by some representational structures (as opposed to purely 'behaviorist' organisms), has object consciousness. Therefore, we could say that object consciousness necessarily goes together with intentionality (if you buy intentionality, you get object consciousness for free): many organisms are object-conscious insofar as they entertain a dynamical sensorimotor relation with the environment. Indeed, object consciousness is a *transitive* form of consciousness: it is always a consciousness *of* (something).

Note that this conception of consciousness entails a clear-cut distinction between consciousness and self-consciousness: one can be conscious of something without being self-conscious but not *vice-versa*. Many animals are conscious without being self-conscious, and the same is true of infants.[11] On the other hand, it is impossible to develop self-consciousness without possessing simple (object) consciousness.

The methodological ground of this distinction comes from a certain way of reading Brentano's theory of intentionality. If ordinary intuition takes consciousness as a phenomenon fully 'internal' to the mind, Brentano

[11] See below, Sect. 3.3. There we will see that there are different types, or degrees, of self-consciousness. Of course, non-human animals and babies do not possess a narrative self.

conceives of it in relational terms: consciousness is not so much a primary and essential quality or character of the mind, but rather a collection of heterogeneous forms of active relations, involving the construction of representations, between an organism and its environment. Against this methodological background, we can speak of an immediate organismic subjectivity, consisting in the primary (object) consciousness of the infant or animal. Primary consciousness is the result of the representational activity. This activity shapes a purely 'objectual' experiential space.

Just to give an idea, our primary object consciousness is similar to Damasio's (1999) notion of core consciousness. Core consciousness is the non-reflective, non-rational experience of the present environment and the body. It concerns solely 'the here and now'. In other words, core consciousness could be defined as the instantaneous or quasi-instantaneous feeling of what happens; it is the happening of those 'mental images' which constitute our contents of raw experience; for instance, a colored surface, a sound and a sensation of heat. However, our position differs crucially from Damasio's inasmuch as we think that being in a conscious state is feeling something without feeling themselves. The next section will be fully devoted to this opposition.

Object consciousness is a bare condition. When infants, almost from the birth, explore the environment, they entertain a rich collection of objectual conscious states. And in exploring the environment they soon discover a particular object: their body. Or, more precisely, they discover parts of their body: they are conscious, for instance, of their hands (without 'knowing', of course, that they are their hands). This is the beginning of an absolutely crucial step, since, in order for an organism to achieve self-consciousness, its consciousness must first apply to a particular object: the body. Indeed, we argue that *the most elementary form of self-consciousness* (*and of subjectivity in the phenomenological sense*) *is the representation of one's own body taken as a whole*.

Here is the point at which we part company with other theorists committed, at least with regard to certain aspects, to the naturalistic and bottom-up strategy. Indeed, there are many authors—just to name a few: Bermúdez (1998, 2007, 2009), Kriegel (2008), de Vignemont (2007) and the already mentioned Damasio (1999, 2010)—who take consciousness of one's own body as the basic component of the self. We agree with these authors at least on the following points: the continuity between

non-human animals and human beings, the existence of a precocious form of consciousness, the centrality of the bodily representation. The main point of disagreement (but a crucial one) concerns the relation between consciousness and self-consciousness. In the next section, we discuss this opposition; this will allow us, at the same time, to present our own view of bodily self-consciousness as the first step in the construction of the self.

3.3 Consciousness and Self-Consciousness: The Case Against Pre-Reflective Self-Consciousness

As we have just said, our feet are firmly planted in the camp of those researchers who take the body as the ground of the notion of self. The first form of personal identity is a bodily identity, that is, the awareness of possessing a body; higher forms of self-consciousness develop from this bodily form of self-consciousness.

However, there are different ways of developing this body-grounded notion of self. In particular, there is a currently highly influential view on which rudimentary forms of self are already present in the newborn. These authors claim that there is a *pre-reflective self-consciousness*, which is an intrinsic component of many, or perhaps all (depending on the authors), conscious states.

This thesis comes indeed in two distinct main versions, which we will label 'inflationary' and 'deflationary' (following Bermúdez, 2011). According to the inflationary version, pre-reflective self-consciousness is a phenomenally salient sense of ownership of one's own mental states that goes along with *any* conscious state, being already present at birth. The most influential authors supporting this view are Shaun Gallagher and Dan Zahavi. By contrast, on the deflationary view, pre-reflective self-consciousness appears from the age of 4–5 months and is implicit to our bodily abilities, *not* being (at least initially) phenomenally salient.

In this section, we argue that the notion of pre-reflective self-consciousness is misleading, especially in the inflationary version, and that talking about self-consciousness does not make sense before the age of about 20 months.

* * *

Let us begin by providing a characterization of the concept of pre-reflective self-consciousness. A good starting point is Gallagher and Zahavi's (2015) inflationary account. According to these two authors, pre-reflective self-consciousness, far from being a second-order mental state whose object is an experience or any other first-order state of mind, is a structural feature of the experience itself; it does not require attention but is tacit and altogether non-observational. As they put it:

> In the most basic sense of the term, self-consciousness is not something that comes about the moment one attentively inspects or reflectively introspects one's experiences, or recognizes one's specular image in the mirror, or refers to oneself with the use of the first-person pronoun, or constructs a self-narrative. Rather, these different kinds of self-consciousness are to be distinguished from the pre-reflective self-consciousness which is present whenever I am living through or undergoing an experience […]. (Gallagher & Zahavi, 2015)

Hence, on this view, pre-reflective self-consciousness is best understood as a necessary component of any experiential state. Yet, *what* is present whenever I am entertaining an experiential state? What, exactly, does pre-reflective self-consciousness consist in? It is a sense of ownership or 'mineness' that is always associated to experience; this sense is such that any of my experiences is immediately given to me as mine, as something belonging to myself. Pre-reflective self-consciousness is not to be regarded as a *relation* between the subject and his experience. It is rather something internal and intrinsic to the experience itself. In other words, this kind of self-consciousness does not have an intentional structure; it is not a kind of *objectual*, or *transitive*, consciousness (= consciousness *of* something), despite being systematically associated to states of objectual, phenomenal consciousness. It is a property of conscious states that makes them *first-personal* mental states, conferring upon them their characteristically *subjective* character. Thus, pre-reflective self-consciousness is what makes experience something belonging to a *subject*, without being either an experience the subject has of himself as a subject—indeed this would be an objectivation of the subject, clearly not available to an infant—or an experience of the experience.

The above-mentioned features of this alleged kind of self-consciousness allow pre-reflective self-consciousness to be already present in the newborn baby: it is a tacit or non-explicit feature which requires no language, no concepts, no reflection, no attention and no meta-representational abilities. This implies what is sometimes called the 'Ubiquity Thesis' (Kapitan, 1999), for self-consciousness is ascribed to a very large collection of (arguably all) conscious states.

What is the evidence for the ubiquity thesis? It is worth pointing out that the existence of pre-reflective self-consciousness is taken to be an empirical claim, which is part of a naturalistic research program. Indeed, Gallagher and Zahavi mention different kinds of (empirical) evidence: phenomenological and psychological (behavioral). However, as we will show, most of this evidence is not compelling, and their account is heavily dependent on a priori assumptions, to the point that it could be labeled as crypto-transcendentalist, or proto-idealist.

Let us start with the phenomenological evidence. Most friends of pre-reflective self-consciousness take for granted that whenever we undergo an experience, we experience it as *ours*. A correct description of our conscious life—Zahavi claims—includes the sense of *mineness* (or *for-me-ness*). The problem is that this alleged intrinsic quality of experience is quite vague, as Zahavi himself acknowledges when he qualifies it as a 'subtle background presence' (2005, p. 124). It certainly sounds obvious that the experience I am living through is *my* experience, but it is far less manifest that this is something I perceive or feel or somehow experience in the experience itself.[12] Moreover, mineness is taken to be already present in infants, but we are not in a position to say that infants have the phenomenology of mineness, and it is not clear how we could ascertain it.

As Schear (2009) suggests, the most promising way to bring out pre-reflective self-consciousness is by using a 'contrastive' strategy, that is, comparing ordinary conscious states with cases of conscious experience that *prima facie* lack any kind of self-consciousness. Examples of these

[12] As Kriegel puts it, '…we may ask, Is for-me-ness one more phenomenal item, or merely a non-phenomenal precondition for phenomenality? That is, is there a *phenomenology of self-awareness*? A deflationist might hold that this for-me-ness is but a dispositional or functional property of conscious states, for example, their global availability to executive function modules; or that it is simply an artifact of the fact that conscious experiences must be someone's experiences' (2007, p. 120).

could be meditative trance or high-level athletic performance. In these kinds of mental states, we are completely immersed in a certain task, and forgetful, so to speak, of ourselves. We are one and the same thing with a certain thing or task. However, this strategy is not available to Gallagher and Zahavi, since it implies that there are conscious but non-self-conscious states, whereas, according to these authors and their followers, mineness is a *necessary* ingredient of consciousness.

We are, therefore, faced with the hard problem of proving the phenomenal reality of a feature which does not clearly manifest itself in experience. Where should we look for pre-reflective self-consciousness as a first person mode of presentation? When we say that mineness is revealed by a 'correct phenomenological description' of experience (Gallagher & Zahavi, 2015, Sect. 2), what are the correctness criteria? If someone were to complain that no sense of mineness was present in his experience, how could he be disproved?

These questions are pressing because, if one lacks an answer and yet insists that the sense of mineness is phenomenally real, he appears to have left the field of a phenomenological psychology and switched to transcendental phenomenology, but this is not the kind of strategy we, or Gallagher and Zahavi, would like to pursue. Nevertheless, we sometimes get the impression that this switch is exactly what Gallagher and Zahavi carry out, perhaps unintentionally.

Given the extreme difficulty of finding mineness in experience, the existence of pre-reflective self-consciousness seems to rather stem from an analysis of the *concept* of experience, resulting in the *a priori* claim that experience requires an owner. Indeed, the suspicion that there is no clear distinction between transcendental argument and (empirical) phenomenological analysis seems to be reflected in some of Gallagher and Zahavi's passages. For instance, when Zahavi claims that we could be helped by a 'hermeneutical intuition' (the expression is Heidegger's), which intends to 'disclose the non-objectifying and non-theoretical self-understanding of life experience in all of its modifications' (2005, p. 79, quoted from Heidegger). Or, when Gallagher and Zahavi (2008, p. 26) argue that a phenomenological theory of self-consciousness that identifies the core of subjectivity in the subjective character or 'first person givenness' of experience should not aim to give a 'description of idiosyncratic experience', but, rather, try 'capture the

invariant structures of experience'. In fact, it is difficult to understand how the attempt to capture the invariants helps us to distinguish phenomenological analysis from a Kantian-style transcendental investigation on the conditions of the possibility of experience in general.[13]

In a similar vein, Frechette (2013) notes that, according to Gallagher and Zahavi, any theory of experience focusing on the first person givenness of experience, without investigating the transcendental conditions of experience, results in a naïve objectivism. Such a theory disregards the fact that 'objects are constituted, that is, experienced and disclosed in the ways they are, thanks to the ways consciousness is structured' (Gallagher & Zahavi, 2008, p. 24). However, as Frechette rightly points out, this assessment of the contribution of phenomenology to an innovative account of self-consciousness could easily be misleading because it connects two issues that are actually distinct—a description of the first person givenness of experience, and the reflective move of transcendental phenomenology—and presents them as if they were necessarily linked.

Consequently, we suggest that the phenomenological analysis does not provide an empirical reason to believe that a sense of mineness exists, or, better, that there is no phenomenal evidence of the empirical reality of mineness. Gallagher and Zahavi unnecessarily turn an objective fact, the fact that when an organism has an experience, it is *its* experience, into a subjective qualitative matter, the *sense* that the organism is the owner of the relevant experience.

* * *

The alleged psychological evidence for pre-reflective self-consciousness comes, to a large extent, from Meltzoff and Moore's research on neonatal imitation (Meltzoff & Moore, 1995; see also Gallagher & Meltzoff, 1996). In sum, these authors claim that neonatal behavior shows that the infant already possesses the cognitive apparatus necessary to perform elementary imitation, and this is interpreted

[13] See Sect. 3.1.3 above. To be sure, Zahavi distances himself from an anti-naturalist interpretation of phenomenology. He points out, for instance, that '[to] naturalize phenomenology might simply be a question of letting phenomenology engage in a fruitful exchange and collaboration with empirical science' (2009, p. 8). However, it remains to be clarified how naturalized phenomenology can provide evidence for the phenomenally salient sense of mineness: *that* is our problem.

as evidence that the phenomenology of body image and the sort of self-other distinction implied in it exist from birth.

However, Meltzoff and Moore's data admit alternative accounts. For one thing, compelling evidence for newborn imitation is lacking. Jones (2009) has argued that only tongue protrusion is reliably demonstrated in neonates, and there is good evidence that tongue protruding is a common response of newborn infants to a range of arousing stimuli, and that a human instance of tongue protruding is one among these stimuli. Thus, it is likely that newborns' matching of tongue protruding in imitation experiments is not imitation, but rather an expression of the infant's interest in, or arousal by, the display of the same behavior by means of which infants typically express interest or arousal. Moreover, mirror systems are often cited as the neural basis for neonatal imitation,[14] but recent research into the development of these systems suggests that, rather than being present at birth, they develop via a combination of Hebbian learning and experiential canalization (see Farmer & Tsakiris, 2012). Finally, even if one is disposed to acknowledge that the newborn is able to perform some *proto*-imitations, this can hardly be considered as evidence of phenomenological structures present in the goal-oriented imitative behaviors displayed by older children and adults (see Welsh, 2006). Indeed, such a reading of Gallagher and Meltzoff's interpretation seems preposterous. It is in fact not clear that having a proto-image of the body involves a phenomenal correlate describable as a sense of mineness. Rather, the baby is able (or so we suppose for the sake of argument) to perform certain intentional behaviors insofar as she possesses certain mental representations of (parts) of her body. Of course, these representations have a phenomenal correlate but no sense of self is required.

On our view, a sort of 'adultomorphization' of infancy afflicts the interpretation of evidence from developmental psychology, ecological psychology and infant research supporting the claim that there is a form of self-consciousness already in place from birth (see Gallagher, 2005;

[14] For example, although there is no direct evidence for the activation of mirror neurons in neonates, Gallagher proposes the following hypothesis: '…when the neonate sees another person perform a specific motor act, for instance a tongue protrusion, the visual stimulus initiates the firing of the same mirror neurons that are involved in the infant's own performance of that motor act' (Gallagher, 2005, p. 77).

Gallagher & Zahavi, 2015). Historically, adultocentrism comes in two forms. It may be 'discriminating' or 'excluding': the newborn is an animal-like creature lacking any kind of mind and a true interpersonal life. Or it may be a top-down approach to neonatal psychology involving empathic identification or projection: the newborn is then viewed and evaluated not from the standpoint of her world but from the standpoint of the adult (Peterfreund, 1978).

This is an old problem. One example is the hypothesis—originating from Janet and Freud, and then widespread in the psychoanalytic literature—according to which the construction of reality in the child begins with the dispelling of a state of primitive confusion between body and world, subject and object and interiority and exteriority (see Lichtenberg, 1989). This hypothesis presupposes just the pre-existence of what is to be built, that is, takes for granted the presence, though in a primitive and confused way, of some form of self-consciousness. This is a mistake that derives from the intuitive but false assumption that the representation of reality can take form only as an object that differentiates itself from a subject experienced as such—a mistake that must be corrected by making (1) the distinction between representation as a form of object consciousness and self-conscious representation, and (2) the distinction between the subject as a mere *functional center organizing action*, existing in any animal with a brain, and the self-conscious subject.

The methodological moral is that the study of the 0–1-year-old infant's subjectivity should follow the example of the study of animal subjectivity, where cogent evidence can be found of very complex inter-individual behavioral dynamics produced by conscious (but not self-conscious) activities of representation. Animal behavior researchers (and especially primatologists) 'are typically circumspect in their interpretations, limiting their claims to operationalizable terms […] rather than making claims about the nature of the experience that may be involved in an animal's performing a task' (Allen & Trestman, 2015, Sect. 7.4). Recently, cognitive neuroscience has shown how to investigate the 0–1-year-old infant's subjectivity limiting one's claims to operationalizable terms. The groundbreaking study by Kouider et al. (2013) shows that neural markers of consciousness found in adults can be generalized to infant populations (5-, 12-, and 15-month-old infants).

Against this methodological background, our view can be sketched as follows. We can speak about an immediate, organismic subjectivity consisting in the primary consciousness of the infant or animal. Primary consciousness is the result of representational activity. This activity, as stated above, shapes a purely 'objectual' experiential space.[15] The crucial point is that having a proto-representation of the body does not imply the unification of the parts of the body. This idea—the idea that the baby is precociously aware of her body as a *unified* body—carries with it the suggestion that the neonatal mind is able to distinguish, *phenomenally*, between the internal and the external. By contrast, since there is not a unified bodily (experiential) space, this distinction is not available; at this stage of development, the objectual field is identical to the subjective world. We could say that the self exists merely as a psychobiological system.

It is here that a different version of pre-reflective self-consciousness arrives on the scene. We are referring here to those moderate positions according to which, even if phenomenal mineness does not exist, very precocious kinds of representation are still to be regarded as forms of (pre-reflective) self-consciousness. This is exactly the view proposed by Bermúdez, who calls it a 'deflationary' account of mineness (as opposed to the inflationary view proposed by Gallagher and Zahavi, de Vignemont and, arguably, Damasio).

According to Bermúdez, visual proprioception *implicitly* carries information on the distinction between the bodily self and the external world—as he puts it, 'self-specifying information', since the self is, in a way, 'perceived' as that which specifies the limit of the visual field: 'The self appears in perception as the boundary of the visual field, a moveable boundary that is responsive to the will' (Bermúdez, 1998, p. 106). Affordances, visual kinesthesis and bodily invariants all carry self-specifying information. Since this information is intrinsic to the working of the visual system, it is very precociously available to the child. The same can be said of other somatic (or proprioceptive) representations, the perception of bodily properties 'from the inside', such as pain or the sensation of losing one's balance.

[15] As Lyyra puts it, 'Originally, only world is given to the subject' (2009, p. 76). This is the author's formulation of what Fonagy, Gergely, Jurist, and Target (2002) call 'psychic equivalence'.

These data should not, however, be taken as evidence for pre-reflective self-consciousness, for the following reason. The representations involved in precocious perceptual states are representations of single parts of the body, not of the body taken as a whole. And when a baby, say, six or eight months old perceives, say, her hand, she perceives it *as an object among others*, not as a *part* of her body. Indeed, in order to perceive it as a part of her body, she would have to possess the ability to represent her body as a whole, which is not the case; for it is over the course of *the first three years of life* that 'an explicit visuo-spatial representation of one's body progresses from early awareness of individual body parts to representation of the body as a whole in which the body parts together constitute a typical configuration that corresponds to others' bodies' (Brownell, Svetlova, & Nichols, 2012, p. 40). Thus, there are no empirical grounds for assuming that infants under 1 year of age are able to construct a representation of the unity of their own body. Indeed, we have reason to think that, at around one year of age, the child is in the process of bringing together some parts of her body. Prior to this point, her body can be said to be made up of 'close' and 'domestic' objects, but these being part of the world, the kind of agentive and phenomenological relation between the infant and, for example, her thumb, is similar to the relation between that infant and, say, her soft, little pillow, imbued with her smell (see Jervis, 2011, p. 82).

Things are not much different in the case of proprioceptive states such as pain. Of course, pain is not an external object, and the painful experience is qualitatively different from, say, the visual experience of one's own hand, yet the relation between the infant and her pain can be assimilated to a perceptual relation with 'something' located somewhere. She is not conscious of the pain *as something that is part of* '*herself*'; she is simply conscious of the pain, of that bad thing. The infant is absorbed in her pain (to the same degree as she can be absorbed in the tactile or visual exploration of her hand), and does not objectify herself as being in pain. Pain (or perhaps pain-here) is the object of her consciousness; there is no reason to account for the qualitative distinctiveness of pain (or of other feelings) in terms of a sort of internal sense of oneself.

Here again, it is worth recalling: like animals, less than one-year children are conscious merely in the sense of being able to form representations of objects and actions. They can have separate experiences of parts

of themselves, but only in the sense that they have experience—and develop knowledge—of parts which *we* know (and they do not know) are parts of them. In the idiom of phenomenology, we can say that the newborn, like the infant at six months or one year of age, produces a rich subjectivity, but, being immersed in it, cannot objectify it. That is, the infant is an active subject in the sense of being a functional center organizing action, but she cannot 'have' either herself or parts of herself, that is, she does not 'have' herself as an active subject. And when, for instance, the infant's eyes are exploring the environment, she 'is', so to speak, her eyes, but certainly she does not 'have' her eyes, and neither, actually, does she imagine their existence (see Jervis, 2011, p. 81).

At this point, it seems that one could draw the conclusion that pre-reflective self-consciousness does exist after all: it emerges when the baby becomes able to represent her entire body. However, this conclusion is hasty, and still too strong, for the self is not a simple object like any other. The notion of self involves, to say the least, the whole body of the organism *experienced as one's own body*. Indeed, the notion of self (or of self-consciousness) brings with it an aspect of subjectivity that is missed in the objectual representation of one's own body. This is clear when the infant—between the ages of 18 and 24 months—becomes able to recognize her specular image in the mirror (e.g., Courage, Edison, & Howe, 2004; Lewis & Brooks-Gunn, 1979; Nielsen, Dissanayake, & Kashima, 2003).

It is true that for the last forty years the significance of mirror self-recognition as an indicator of self-awareness has not gone unchallenged (see Parker, Mitchell, & Boccia, 1994; Suddendorf & Butler, 2013). Overall, there are *lean* interpretations (e.g., children pass this test because of kinaesthetic-visual matching skills), *rich* interpretations (e.g., children's mark-directed behavior is evidential of an introspective form of self-consciousness and a self-concept inherently linked to understanding the mental states of other people) and proposals lying somewhere between the two. Taking, as we do, mirror recognition as a marker of *bodily* self-consciousness, falls within this last option.

Thus construed, mirror self-recognition involves being able to form a bodily image of oneself as an entire object, and simultaneously to take this image as a *subject*, that is, as an active source of the representation of oneself. Here the subject recognizes a new kind of object of consciousness:

the object is the subject itself, or better the objectified image of the subject—'it is *me* there'. That this marks the agent's achievement of self-objectivation as 'me' is also supported by the evidence that verbal and deictic self-reference and mirror self-recognition develop in close conjunction (see Lewis & Carmody, 2008; Lewis & Ramsay, 2004). Mirror self-recognition onset indicates, therefore, the emergence of a new modality of cognition compared with the ability to build the image of any external object that is characteristic of animal consciousness in general, one that is unique to humans and only a few of the higher non-human primates (see Lewis, 1994).

Thus, at most we could concede that it is possible to distinguish between an objective self (= the whole body of the organism) and the subjective self; yet it is the latter that better fits the ordinary concept of self-consciousness. Referring to the organism as a 'self' is misleading to a certain extent, although the claim that the 'objective self' is a necessary condition for the development of a subjective self is certainly correct. But, to repeat, in order to ascribe a (subjective) self, the representation of the body as a whole must somehow make explicit that the represented body is one's own body, that is, the child must be able to take the representation of his body both as an object and as a subject which is the source of the representation, and this is not the case before the age of (at least) 18 months. Further developmental stages are required.

* * *

Let us take stock.

The deflationary version of pre-reflective self-consciousness antedates the appearance of (bodily) self-consciousness at the age of 4–5 months. But we have seen that before the age of (more or less) 18 months the infant does not possess the relevant kind of representation: s/he lacks the involvement of the representation of the whole body as one's own body.[16]

As to the inflationary version, it is hard to avoid the impression that, in Gallagher and Zahavi's account, pre-reflective self-consciousness has been characterized only *negatively*, that is, as a kind of self-consciousness

[16] At most we are disposed to concede to Bermúdez that there might be some forms of pre-reflective self-consciousness (at the age of about 18 months), but merely in the sense that it is hard to say whether the representation of one's own body as one's own is definitely conceptual. On the non-conceptual versus conceptual character of self-consciousness, see Musholt (2013).

that does not involve any of the capacities/processes associated with a full-fledged self-consciousness. When it is time to offer a positive description of pre-reflective self-consciousness, all we are told is that it is something that figures 'as a subtle background presence' within (the philosopher's) object-consciousness (see Sect. 3.2) or, still more elusively, a phenomenal property that is 'an unstructured intrinsic glow' (Kriegel, 2008, p. 363). Briefly, the notion of pre-reflective self-consciousness resists positive description.

For this reason, it seems to us that Gallagher and Zahavi's proposal may best be assessed as the most recent legacy of a venerable philosophical tradition that goes back to Kant and, through Husserl, arrives at Sartre. It is the tradition in which the relation between consciousness and self-consciousness has been regarded as self-evident and necessary, in force of transcendental-style arguments. Of course, we are not saying that the Kantian notion of self and the phenomenological notion are one and the same thing. As Zahavi rightly points out, for Kant the self is not *given*, whereas in the phenomenological tradition, at least as Zahavi reads it, selfhood is similar to a primary *datum* (see Zahavi, 2005, Chap. 5). In particular, it is clear that Zahavi's 'minimal' or 'core' self is not identical to the Kantian abstract and formal principle of unification, as is evidenced by his frequent reference to the notion of 'immediate givenness'. Nevertheless, the very existence of the self, in Gallagher and Zahavi's neo-phenomenological view, rests upon a sort of *a priori* indisputable evidence quite similar to the conceptual necessity underlying the Kantian *I think*, and, in some cases, the pre-reflective sense of mineness is explicitly supposed to play the unifying role of the *I think*. For example, building on Gallagher and Meltzoff, Rochat (2012) interprets some findings about early development as suggesting that neonates manifest 'unity in the Kantian sense', that is, 'a primordial sense of an embodied self-unity'. Here it is difficult to avoid the impression that this thin, minimal form of self-awareness is the result of a psychologistic hypostatization of the Kantian unity of synthesis—it could be said, as a sort of slogan, that the function of integration of experience is confused with the experience of integration.

Hence, pre-reflective self-consciousness is a sort of reification of a *desideratum* coming from an understandable need: *there seems to be something in our very way of thinking about ourselves that forces us to believe*

that, *whenever we are conscious of something*, *we are*, *at the same time and in a certain way*, *conscious of being so*. In this sense, pre-reflective self-consciousness is an ineluctable obsession, a sort of permanent cramp in our thinking. However, *this I is rather the product of a reflective activity*, and projecting the result of this activity onto the newborn is a form of illegitimate adultocentrism: the phenomenology of selfhood, namely, the feeling of existence, is something *constructed* starting from bodily self-consciousness.

Zahavi dismisses the view, like our own, in which the baby's consciousness is described as a presentation of the world or as an 'immersion' in the world, by arguing that it implies an unacceptable consequence, namely, that 'whatever experiences they [= the babies] have are present to them in a third-person manner, that is, in the same way as publicly available objects are' (2015, p. 147). However, it is not clear why it is unacceptable. As far as we can tell, the only reason he offers is the argument typically brought against first-order representational theories of consciousness, namely, that they are unable to distinguish between conscious and non-conscious states. In fact, in the targeted accounts (such as ours) of the newborn's mental life, conscious mental states are states 'we are conscious with and not states we are conscious of',[17] that is, (phenomenal) consciousness presents us with nothing except for external objects and their properties, and it is unclear to what extent such a view really allows us to distinguish conscious and non-conscious mental states, all of which allegedly represent objects in the environment.

How do we deal with this objection? Our reply consists, basically, in biting the bullet and somewhat reversing the charge: it is true that in some cases we are not in a position to establish whether an organism is in a (phenomenally) conscious state or not; yet this is due to the fact that, in the absence of clear behavioral evidence, the notion of 'conscious state' is intrinsically vague. And this is no greater a problem for us than it is for our opponents.

Take for instance the situation in which someone is looking for a bunch of keys; the keys are right in front of his eyes, but, for a while, he does not notice them—he does not 'see' them. Could we say that he is conscious

[17] To quote Dretske (1995, pp. 100–101), one of the most prominent advocates of first-order representational theories of consciousness.

of the bunch of keys? It is not clear. Dretske's answer is positive. If we accept this answer, we may account for the difference between that state and the state in which he eventually notices the keys in terms of a difference between a simply conscious and a self-conscious state, or in terms of a difference between non-availability and availability of (conscious) information to cognitive systems. But it seems to us equally reasonable to say that the man was initially not conscious of the keys.

Examples like these can easily be multiplied (e.g., Tye, 2003, I.2). The crucial question is: Can the notion of pre-reflective self-consciousness help us in such cases? We cannot see how. To say that mineness is present in one case but not in another seems preposterous. Let us put the things in the following way. According to a rough definition, in order for a mental state to be conscious, the information carried by that state must be available to the organism as a whole. What is added to this rough definition by the suggestion that it is mineness that distinguishes conscious states from non-conscious states? Is it not, after all, just another name for what we referred to above as 'availability to the organism as a whole'? The idea of mineness merely creates the illusion of having a criterion to discriminate personal (conscious) from subpersonal states, but it is just an idle wheel, since we no longer possess a criterion to check whether mineness is present. Moreover, Gallagher and Zahavi's account lacks the resources to distinguish between consciousness and self-consciousness, and this, as seen above, has negative consequences, certainly worse than the real difficulty of providing a criterion for discriminating conscious states from subpersonal intentional states.

To put it briefly, Gallagher and Zahavi start from a fully 'personal' view of *phenomenal* consciousness, which somewhat begs the question of the existence of pre-reflective self-consciousness (in favor of it). Indeed, they argue that self-awareness is related to the idea that experiences have a subjective 'feel' to them, a certain (phenomenal) quality of 'what it is like' or what it 'feels' like to have them: 'what-it-is-like-ness is properly speaking what-it-is-like-*for-me*-ness' (Zahavi, 2015, p. 145). But this cannot be taken for granted—even within the framework of phenomenology. For one thing, as Schear (2009) rightly notes, such an emphasis on phenomenal consciousness risks neglecting the existential dimension of the problem of subjectivity in the phenomenological tradition: 'In what

sense could phenomenal consciousness constitutively involve the possibility of bad faith or existential disorientation?' (p. 104). This is a remarkable point since bad faith and existential disorientation will be significant themes in the next two chapters.

Secondly, being a phenomenologist did not prevent Merleau-Ponty from entertaining the possibility of an awareness which is devoid of selfhood, a 'consciousness that is neither self nor other' (Merleau-Ponty, 2010, p. 29; for a commentary, see Welsh, 2007, 2013).

3.4 The I as the Making of the Me

Pre-reflective self-consciousness is also at the core of Prebble, Addis, and Tippett's (2013) theoretical framework for investigating autobiographical memory and sense of self. In this framework, pre-reflective self-consciousness is seen, inter alia, as the key to understanding the Jamesian notion of the 'I' (the self as knower) as opposed to the 'Me' (the self as known). We will now argue that this reading of James' notion rests on a serious misunderstanding of his theory of the duplex self. This is rather important since James' distinction is ubiquitous in the theoretical discussion around sense of self.

Prebble, Addis and Tippett's model of the sense of self hinges on two features varying along two axes or dimensions. On the first axis we find the opposition between a subjective and an objective aspect of the sense of self, where the former is our conscious, phenomenological experience of selfhood ('subjective sense of self'), and the latter our mental representation of self, comprising all the things we perceive and know about ourselves ('content of self'). On the second axis we find the opposition between those aspects of the sense of self that are related to the present moment ('present self') and those extended over time ('temporally extended self-concept'). In each particular moment I experience a sense of (synchronic) unity both in my conscious experience of selfhood ('subjective sense of self') and in my mental representation of who I am ('self-concept'). Over time, I experience unity both in my subjective experience of selfhood ('phenomenological continuity') and in the way I mentally represent myself across time ('semantic continuity'). Thus, we get four components of the sense of self, the simpler functions being necessary precursors for more complex functions.

Prebble, Addis and Tippett's model distinguishes two hierarchically related forms of present-moment conscious self-experience: 'pre-reflective self-experience' and 'self-awareness'. These are both necessary (though not sufficient) for autonoetic consciousness (autonoetic recollection and autonoetic imagining) and episodic memory, but pre-reflective self-experience both precedes and grounds self-awareness.

Now let us compare this theoretical framework with James' theory of the self.

In Chap. 10 of *Principles of Psychology* James begins by noting that both the common man and the spiritualist philosopher are spontaneously led to suppose that in one's own experiential space there is an innermost center, a dynamic center of initiative and free will ('the active element in all consciousness') denoted by the pronoun 'I'.[18] James defines it as 'pure Ego', and notes that philosophers' interpretations of it lie along a spectrum that includes, at one end, the claim that it is 'a simple active substance, the soul', which is the metaphysical guarantee of the presence of the self to the world, and at the other, a Humean perspective claiming that 'it is nothing but a fiction, the imaginary being denoted by the pronoun I' (James, 1950, p. 298). James' theory of the self, however, is more complex: it first takes the Humean step, but then goes beyond it.

If I say, 'I picked up the book from the table', the pronoun 'I' refers to me as an *agent organism*, taken as a whole and as opposed to an external object. In this case, the book is a *completely* external object, but sometimes I (as a global agent subject) can also consider an object that is not entirely external, such as my foot (that is part of my being but nevertheless 'down there'), or my hand, or even something else that is more 'here' (or 'less there') than my foot—for instance, my eyes or my head, which are *almost* part of the intimacy of the ego. In all these cases I keep detaching and differentiating my subjective ego, as a primary psychic subject, from all these other things, which are objects for the ego. Thus far, therefore, I am still rather certain of what my subjective ego is. But then, like

[18] We have here a much thicker intuition than Zahavi's. It includes, in addition to the primary ownership of the self, a sense of agency: in establishing any sort of active relationship with the world, the individual feels that she is moving from a center in her inner space. In this case the presence is not in the background; rather, it is a starting point, a base upon which the individual proudly sets her foot when she thinks she can say 'I'.

anyone, I realize that I am also able to consider as objects things that are much more 'inner'—for example, the global image of my body, a sensation, a smell, a dream, a thought or a mood, such as anxiety or euphoria. I realize then that there is no way to stop this 'hemorrhaging' of my ego: in introspectively probing my mind, I keep taking as an object anything it contains, thus detaching it from myself. What I ask, then, is the following: if all these aspects of the mind are objects—insofar as they are objects of my introspective consciousness—what is the real subject, that is, the wellspring of consciousness? In other words, how can I capture the conscious subject who introspects, if any aspect of myself that I introspectively grasp is only an object of this supposed conscious subject? The innermost ego, as center and driving force of any possible subjectivity, ends up being a pure grammatical trick, a sort of dimensionless point—or, more unsettlingly, the 'wavering and unstable phantom' evoked by Schopenhauer in a famous passage (Schopenhauer, 1969, vol. 1, p. 278n.). Ultimately, this subjectivity is a convention; it cannot be located anywhere. The subject, taken to its limit, does not exist.

Thus far the Humean *pars destruens*, but James does not stop here. Once the agent and observing self has melted into an abstract and depthless subjectivity, James grounds the existential feeling of presence in the subject's experiencing itself as the *empirical self* (the Me-self). This is the way one presents oneself to oneself, thus objectifying oneself in the introspective consciousness of oneself. This self-presentation is a *description of identity*, which—as mentioned in Sect. 2.4—comes in three forms of reflexive experience: the material, social and spiritual selves.

We interpret James, then, as arguing that the I-self is a process of objectivation, which produces the Me-self. The I-self is not 'a metaphysical entity that stands outside our stream of consciousness as the subject of our experiences.' It is not even an implicit, pre-reflective self-awareness, 'understood as an integral feature of our conscious experience of the world' (Prebble et al., 2013, p. 821). The I-self is rather a *process*, the self-representing of a psychobiological system.

One implication is that there cannot be a 'subjective sense of self' (not even a 'brute' first-personal experience) without a 'content of self'; our conscious, phenomenological experience of selfhood *is* our feeling of being here as being here *in a certain way*, according to a mental representation

'comprising all the things that we perceive and know about ourselves' (ibid., p. 817). This is well captured by Dan McAdams: the I-self, he makes clear, 'is really more like a verb; it might be called 'selfing' or 'I-ing', the fundamental process of making a self out of experience' (1996, p. 302). The Me-self is instead 'the primary product of the selfing process;' it is 'the self that selfing makes' (ibid.). The Me exists as an evolving collection of self-attributions (James' material, social and psychological selves) that result from the selfing process. It is 'the making of the Me that constitutes what the I fundamentally is' (McAdams & Cox, 2010, p. 162).

So construed, James' theory of the duplex self entails that there is no consciousness of self without *knowledge* of self; I know that I exist insofar as I know that I exist *in a certain way*, that is, with particular features, as a describable identity. And it is to be noticed that this claim contradicts what Kant asserts in a famous passage of the first *Critique*. As is well known, Kant agrees with Hume (and James): the empirical apperception 'can give us no constant or enduring self in the flow of inner appearances' (Kant, 1998, p. 232, A 107). Yet, he thinks that one may shift from the analysis level of psychological experience to that of transcendental arguing, and here posits a *pure* apperception: 'I am conscious of myself, not as I appear to myself, nor as I am in myself, but only that I am', he writes in the first *Critique* (B157); and in B158 he adds that '[t]he consciousness of self (*Bewußtsein seiner selbst*) is […] far from being a knowledge of the self (*Erkenntnis seiner selbst*)'—that is, the consciousness of *existing* is distinguished from the consciousness of *existing in a certain way*. Thus Kant's I think ('that accompanies *all* my representations') is something undetermined and void ('a something = X'), which, not unlike Descartes' cogito and Zahavi's subtle background presence, lays a claim to being a *primum*.

Things look very different from the standpoint of Brentano's relational conception of consciousness outlined in Sect. 3.2. Here the conscious mind is seen as a set of heterogeneous forms of active relationship between a living organism and its world-environment. We are not conscious in the abstract, but we are always conscious *of* something, and, among all the representations of object, there is a representation concerning the subject itself, and this is self-consciousness. In this framework, self-consciousness is no longer a primary, elemental, simple awareness of the self, preceding any other form of knowing; rather, it is a variation of our relationship to the world.

Thus, as Schopenhauer had already noted, and unlike Kant, Brentano thinks that self-consciousness is not a basic modality of consciousness, is not a primary and simple 'knowing of being-there', but 'consists in watching oneself, seeking after oneself, and hence it is from the very beginning a knowing of *being-there in a certain way*' (Jervis, 2011, p. 71). Indeed, Schopenhauer had already considered the possibility that 'this knowing of being-there is never exhaustive, in the sense that it is a search for itself always unsatisfactory, and hence interminable' (ibid.).

This way of conceiving self-consciousness is the fundamental premise of our reconstruction of the pathway through which, starting from the awareness of ourselves as bodily agents, we become aware of our mental life. There is no consciousness of self without knowledge of self: I know that I exist insofar as I know that I exist in a certain way, that is, with particular features, as a describable identity. The notions of self-consciousness and identity cannot be separated.

Thus self-consciousness is a self-describing, an identity forming. It is a unifying, integrative, synthesizing process (see McAdams, 1997, p. 56), a synthetic function, although not a Kantian one. Kant's arguing that the synthetic unity of apperception is the transcendental condition for having any representational state builds upon the picture of a subject who is originally unitary. In Kant the person is always given in its unity, as if the psychological level of analysis was always and in any case guaranteed by the logical/transcendental level of analysis. This, however, does not apply to the synthesizing selfing process: as we will argue in Chap. 5, the empirical subject is primarily non-unitary and gains its unity in the act of mobilizing resources against the threat of disgregation. In this perspective, the unity of apperception is a process and an achievement.

4

Making the Self, II: Psychological Self-Consciousness

In Sect. 3.4, we argued that being self-conscious consists in knowing that one exists as a describable identity. In this perspective, the development of self-consciousness is the process of construction of different forms of self-identity.

The earliest form of self-identity is a *bodily self-image*. In Sect. 3.3, we saw that in the course of the second year the child succeeds in the complicated operation of forming a bodily image of herself as an entire *object*, and simultaneously taking this image as a *subject*, that is, as an active source of the representation of herself. This acquired awareness of the body as one's own is the basic premise necessary to provide ourselves with that elementary reflexivity that allows us to know that we exist. Thus self-consciousness in its most basic form, namely as awareness of one's own existence, is the perception of a physical identity. Once more: it rests not on a supposed pure and primary feeling of existing, but on a *self-describability*—the child gains access to the feeling of existing when she recognizes herself in a body distinguishable from others' bodies, when she comes to know herself as a bearer of physical, physiognomic bodily features.

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_4

The child who begins to master the subjective-objective space of the body, however, will need to take the further step of appropriating of the virtual inner space of the mind. That is, she will need to be able to objectify her own subjectivity, knowing that it is *her own* subjectivity, in the same way in which she is able to objectify her own body, knowing that it is *her own* body.

The construction of the virtual inner space of the mind is the topic of this chapter, which will work back and forth between theoretical psychology and the findings of empirical research. Within the framework of attachment theory, we will draw on developmental, social and personality psychology to reconstruct the process through which, starting from bodily self-awareness, we become aware of the existence of the mind as a virtual inner dimension. This awareness is a psychological form of self-consciousness that will evolve into the most cognitively demanding form of self: a narrative (or autobiographical) self. We thus part company with all those accounts of narrative identity that pay little or no attention to the role of the body in the development of the narrative self-concept; we will see that without an affective and bodily self-description, narrative selfhood would not arise. On the other hand, as we saw in the previous chapter, our account rejects the hypothesis that the embodiment of the narrative self is provided by a pre-reflective self-consciousness viewed as a primitive, proprioceptive form of self-consciousness already in place from birth.

The agenda of the chapter is the following. In Sect. 4.1, we probe the nature of introspective consciousness. In Chap. 2, we introduced Freud's idea of a pervasive presence of self-deception in our inner life. This critical approach to introspection has found a rich source of evidence in the cognitive neuropsychology of confabulation and in the social psychology literature on cognitive dissonance and self-attribution. What these research traditions deliver is a drastically anti-Cartesian picture of introspective consciousness; where Descartes saw a given essence (the self-transparent consciousness-substance), there is now something *constructed* (the Humean 'theater'), which is the product of an apparatus that allows us to partially describe, and above all narratively justify, mental processes all of which are fundamentally unconscious.

In Sect. 4.2, the focus is on the ontogenesis of the virtual, inner 'theater' of the mind. It will be argued that the construction of introspective experiential space occurs through the process of turning one's mentalistic skills on oneself under the communicative pressure of micro-social contexts. We will first look at *affective* mentalization, arguing that a good attunement in the proto-conversational infant-caregiver interactions plays a crucial causal role in the construction of the phenomenology of basic emotions. We will then examine how the construction of an inner experiential space advances under the thrust of caregivers' mind-minded talk.

The focus of Sect. 4.3 is on the emergence of a self which is continuous through time. Here we interpret the claim that a minimal self is a precondition for an autobiographical self as the claim that the most minimal form of self-consciousness is a nonverbal, analogical representation of the bodily self that acts as a fixed referent around which autobiographical memories can start being organized.

In a nutshell, the psychological evolves from physical self-description through an interplay of mentalizing abilities, autobiographical memory and socio-communicative skills. Section 4.4 makes it clear that this interplay is modulated by sociocultural variables. Data from cultural psychology show that introspective consciousness is not an all-or-nothing phenomenon. In normal adults in pre-agricultural or pre-literate agricultural cultures the incompleteness of the capacity to conceptualize the existence of an inner experiential space can be observed.

Lastly, Sect. 4.5 comes to grips with narrative identity. Following Dan McAdams' well established theoretical systematization in the field of personality and personality development, narrative identity is defined as the ability to construct 'an internalized and evolving story of the self' that can provide life with 'some semblance of unity, purpose, and meaning' (McAdams & Olson, 2010, p. 527). Although our theory incorporates certain aspects of a narrative approach to the self, we are careful to distance ourselves from the hermeneutical versions of narrativism. We argue that the active process of self-interpretation that is constitutive of personal identity is a theory-driven narrative re-appropriation of the products of the neurocognitive unconscious. Thus the self is a self-interpreting being in a naturalistic sense—a sense which has thus far been foreign to the hermeneutical tradition.

4.1 The Nature of Introspection

The concept of self-consciousness is twofold, involving two different levels of complexity. The (relatively) simpler level is that examined in the previous chapter: self-consciousness is consciousness of the self as representation of the unity of one's own body, which is an experiential space that is singular and ambiguous, neither 'inner' nor 'outer', but simultaneously source and object of the representation. At the more advanced level, on the other hand, self-consciousness is something more complex: it is the introspective recognition of the presence of the virtual inner space of the mind, separated from the other two primary experiential spaces, that is, the corporeal and extracorporeal spaces. We have here the foundation of human self-consciousness in the Lockean sense of the term: self-consciousness as *identity of person*.

According to Locke, the concept of person is not an essence but rather a psychosocial attribute that is assigned to those subjects who possess a specific set of psychological capacities. This is in agreement with the most common legal language, which suitably speaks about 'natural persons' and similarly about 'legal persons', thus pointing out something precise, that is, the presence of an agent or subject who, in virtue of her intrinsic characteristics, is fully able to perform such acts as buying real estate, making a donation or a will, or paying taxes. Here the acting subject is a person precisely to the extent that she can be held (ethically even before legally) responsible for what she does. And she is thus imputable as well; if she committed a crime, she knew very well what she was doing. The concept of person therefore rests on that of *personal responsibility*; it is easy to see, even intuitively, that the concept of responsibility rests on the concept of consciousness, or better self-consciousness, seen precisely as awareness of one's own acts, and hence as *critical appropriation* of one's own projects, actions and memories. An individual can make a will only if she is a person—and indeed a child cannot make a will, or even an elderly person who suffers from dementia; they are not sufficiently responsible inasmuch as they are not sufficiently aware of the meaning, scope and consequences of their actions.

Thus, as already mentioned, the Lockean person is someone who possesses a set of psychological capacities. It is someone who is able to form imaginary test scenarios in order to make a planning evaluation of what can happen as a consequence of his actions. But above all it is someone who is able to grasp himself not only as a material agent in his own present, past and future acts as 'public' acts, but also as an entity who has an interiority, that is, an inner experiential space in which thoughts and affects can be situated as 'private' events. Only someone with sufficient access to her own interiority (to herself as objectivated in the introspective consciousness of the self) can appropriate 'Actions and their Merits' (Locke, 1975, p. 346).

In Locke, therefore, an individual is a person only insofar as she can reflectively appropriate her actions and their meaning—an appropriation that originates from 'that consciousness which is inseparable from thinking' (ibid., p. 335). Locke also realizes that the identity of persons resides in 'sameness of consciousness' rather than sameness of substance: 'So that self is not determined by identity or diversity of substance, which it cannot be sure of, but only by identity of consciousness', he writes (ibid., p. 345). What is truly new in this philosopher is that for the first time consciousness is a 'secular' notion; it is not an innate substance, and above all it breaks with the soul. But if the identity of persons is determined by consciousness, by what is consciousness determined?

Locke relies on introspective consciousness as the most psychological and less metaphysical notion he can conceive to define the concepts of person and identity. On closer view, however, this consciousness is a 'strong' stand-in for the soul; it is, actually, still a sort of secularized soul. Despite the philosopher's good intentions, it is also described as a sort of essence. For all that, Locke's consciousness is still given a priori: it is not something that is constructed during life, which emerges from the multifarious qualities of the body and of human existence. Such a notion of consciousness is found instead in the psychological sciences. As we will see in Sect. 4.3, cognitive scientists have developed Locke's insights, broadening his 'notion of sameness of consciousness into a more general notion of psychological continuity and defending the suggestion that it can define the persistence of persons' (Schechtman, 2013, p. 453).

4.1.1 Being Able to Say Why

When introspection is put under the magnifying lens of the cognitive sciences, however, the question arises as to whether it represents a direct access to mental life, or rather a form of self-interpretation.

Both the classical empiricist and the classical rationalist pictures of introspective self-knowledge have granted it a special epistemic authority. The subject is conceived as transparent to itself, and the reflective awareness the mind has of its own contents is supposed to provide knowledge enjoying a special kind of certainty, in contrast with our knowledge of the physical world.

With the rise of a science of the unconscious, most philosophers took a substantial step back from the claim of the self-transparency of the mind. And the current philosophical and psychological debate on introspection displays a range of positions that vary depending on the more or less radical attitudes toward the implications of cognitive-science work on the subject (see Engelbert & Carruthers, 2010; Schwitzgebel, 2014). At one end of the spectrum lie theories that preserve little or nothing of what has traditionally been ascribed to introspection, and offer *non-introspective* accounts of self-knowledge. At the other extreme are those theorists who continue to believe that the access to at least *some* mental events (e.g., some of one's own thoughts) is different in kind from the access to other people's mental events.

One field in which these different approaches to introspection have confronted each other is the so-called 'Theory of Mind', the area of cognitive science that investigates the nature, ontogeny and phylogeny of our mentalistic capacities. These are the skills that enable us to treat agents as the bearers of unobservable psychological states and processes, and to anticipate and explain their behavior in terms of such states and processes. During the 1980s and 1990s most of the work in this area was concerned with the mechanisms that subserve *third-person* mentalization (henceforth 'mindreading'); but in the last decade an increasing number of psychologists and philosophers have proposed accounts of the mechanisms underlying *first-person* mentalization (or introspection). This new stream of research has required a synergy with other research traditions, most notably studies on confabulation about motives in cognitive neuropsychology and social psychology (see Sect. 2.2).

The experimental study of confabulation has its roots in the initial observations regarding post-hypnotic suggestions in the late nineteenth century, which played a significant role in the birth of psychoanalysis. A deeply hypnotized subject may be instructed to carry out some action (e.g., to walk around the room three times) in response to a specified cue subsequent to the termination of hypnosis. When the subjects emerge from the hypnotic trance, they not only carry out the instruction they received while hypnotized, but will often confabulate a plausible explanation for their action in terms of free personal choices.[1]

Similar phenomena came to light almost by chance in Wilder Penfield's brain stimulation research in the 1950s. If in a patient who is conscious and has been locally anesthetized the subcortical structures that subserve the execution of complex movements are stimulated by means of electrodes, she will execute the movements quite automatically, but if the surgeon asks the patient to say why she moved in that way, she will have no trouble making up explanations in voluntary and rational terms, with no inkling of their falsity: 'When asked "What are you doing?" the answers were, "I am looking for my slippers", "I heard a noise", "I am restless", and "I was looking under the bed"' (Delgado, 1969, pp. 115–116).

Since the 1970s, the theme has been investigated more systematically. A now classic piece of cognitivist research concerns the effects of subliminal or 'suppressed' stimuli during experiments on selective attention. One technique is that of dichotic listening. If subjects are made to listen to two different messages on a tape through separate earpieces, at first they experience great confusion without any understanding, but inevitably end up attending to only one of the messages, suppressing any awareness of the content of the unattended message. It can be shown, however, that the suppressed message has been processed and can influence the subject's thoughts and behavior. For example, if the suppressed message contains instructions concerning how to interpret the message of the attended ear (in cases in which the latter message is ambiguous) the subjects will follow them. But in this case too, they will confabulate a plausible explanation for their choice of a particular interpretation of the attended message (see Greenwald, 1992; Lackner & Garrett, 1973).

[1] For some historical background, see Wegner (2002). For a recent sample of the literature on post-hypnotic confabulations, see Wheatley and Haidt (2005).

However, confabulation has been still more systematically investigated in cognitive neuropsychology (especially the clinical-experimental cases of split-brain patients) and in the extensive literature on cognitive dissonance and causal attribution that has been built up in experimental social psychology over the last 50 years (see Carruthers, 2011, Chap. 7; Nisbett & Ross, 1980; Nisbett & Wilson, 1977; Schwitzgebel, 2014, Sect. 4.2; Wegner, 2002; Wilson, 2002).

These research traditions have delivered an enormous number of experiments showing a mismatch between the explanatory *motives* that subjects report to account for their behavior and the *motivations* (i.e., the multiple real causes) of their behavior.[2] In other words, in these experiments the participants have no direct access to the real causes of their behavior; rather, they engage in rationalization or confabulation, that is, they make use of socially shared explanatory theories or of an idiosyncratic theorizing, to fabricate reasonable but imaginary explanations of the motivational factors of their behavior. As Nisbett and Wilson famously put it: 'Subjective reports about higher mental processes are sometimes correct, but even the instances of correct report are not due to direct introspective awareness. Instead, they are due to the incidentally correct employment of *a priori* causal theories' (1977, p. 233).

This is a skeptical and iconoclastic way of looking at human behavior which lies at the intersection of various traditions of thought including, on the one hand, a tradition of critical thought that refers not only to Freud's notion of rationalization but also Marx's concept of ideology, and on the other, the objections raised both by the behaviorists and by Wittgenstein and Ryle against the naïve mentalism. Within this framework, the naïve mentalistic assumption that there must always be a mental event (an intention) that precedes and determines any single action turns out to be the root of an ideology which Ryle (2009) defined as 'the intellectualist legend'. Intelligent behavior, the philosopher argues, despite being the product of indescribable 'know-how', is normally treated as if it were built on the basis of 'intellectual', and as such describable, knowledge.

[2] The critical study of 'motives', regarded as the interpretations of conduct which are 'in force' and 'acceptable' in accordance with the conventions of a specific cultural context, can be traced to Wright Mills (1940), and has been pursued mainly in sociology, as part of the development of symbolic interactionism and ethnomethodology.

This describability ('knowing how to explain one's motives', 'knowing why'), however, is *a posteriori*, that is rather than a planning character, it has a *justificatory* one.[3]

Let us suppose, then, that someone asks me: 'Why are you here?' If I find myself in a certain place at a certain time, it is very unlikely that I can identify the enormously complex chain of motivational factors that have led me to be in that specific place at that precise time. But I will certainly have no hesitation in providing convincing explanations to justify my actions. In short, people can seldom say why they are in a specific place, but can always assert that it is *right* for them to be in such a place (see Jervis, 2007, pp. 156–157).

It is thus that the deceptive character of the folk-psychological inclination to 'read' *any* behavior as deliberately, consciously goal-directed, in accordance with an intention that we assume to be simple and identifiable, comes to light. Actually, the agent is not a primarily *quiescent* organism, who 'then' invariably moves toward some goal; an agent, rather, is a primarily *self-propelled* structure. It is, therefore, improper to ask when a given action was initiated; or even when a given goal-directed behavioral plan began to take shape inside us. It is more correct to say that we have always been embedded in a system of cognitive-motor schemes (knowing how to walk, knowing a language, knowing how to recognize faces and expressions, and so on) and social know-how (scripts) which we began to articulate when we began to exist as individuals, and which we restlessly modify and repropose according to the circumstances. We have always been in motion toward something: in a sense, there is no really new initiative within this flow, because any passage of our life not only occurs under the thrust and in the fluidity of a complex stream of interactions, but also reuses (re-adapts) structures already developed sometime in the past. And, embedded in this flow of actions, we make appeal to declarative or descriptive knowledge to say, or better to tell ourselves: 'This is just the thing I want to do', or 'What I did is the thing that I really wanted to do', and again 'This thought is just what I feel like thinking'. In a word,

[3] According to Tilly (2006), given reasons fall into four overlapping categories: *conventions*, that is, stereotypical explanations; *stories*, that is, explanatory narratives; *codes*, for example, legal and religious formulas; *technical accounts*.

we consider a behavior as deliberate and voluntary any time we are able to describe it within the framework of a coherent and socially admissible rationalization (see Marraffa, 2011b, pp. 181–182).

This explains why there are cases (e.g., cases of akrasia) in which the usual mechanisms underlying the descriptive re-appropriation of our behavior and mental work fail. Failure in this sense occurs when we attempt to superimpose some extraneous and incongruous interpretative pattern on our actual behavior—a pattern that is not relevant to the way in which we actually organize actions and emotions in the flow of our real life. Thus, in an overturning of the traditional approach to theory of action, we are asking not how behaviors that contradict our intention can exist but, on the contrary, if ever deliberate and voluntary behavior ever exists—just as in Chap. 2, when we started by asking how consciousness, rather than the unconscious, is possible.

Here a caveat is in order. Ryle's criticism of an over-intellectualistic picture of human agency (on which we are building) should be kept well separated from his rejection of propositional attitude realism, namely the claim of the existence of mental entities that are causally efficacious, content-bearing, physically realized internal states. Thus, we assume that his insights can be put to work within the framework of a revisionary approach to intentional psychology which is, however, not eliminative—in keeping with our approach to the interface problem in Sect. 2.3.2.

To sum up, introspection, construed as a source of knowledge of the multifactorial etiology of our judgments, decisions and behavior, is an illusion. In its stead we find a mechanism that subserves an activity of *a posteriori* descriptive re-appropriation of the outputs of the cognitive unconscious's processing—our capacity to explain our judgments, decisions and behavior *ex post* as the products of a rational and autonomous agent. In most cases of everyday life, then, giving reasons for what has been done (being able to say why) plays a justificatory rather than descriptive role.

4.1.2 Self/Other Parity or Inner Sense?

Within the theoretical and experimental framework we have been outlining thus far, agents enjoy no introspective self-knowledge of the causes of their behavior; rather, they are engaged in an *interpretative* activity that depends on mechanisms capitalizing on explanatory theories that apply to the same extent to themselves and other people. Such mechanisms are triggered by information about mind-external states of affairs, that is, the subject's behavior and the situation in which it occurs—information, therefore, with respect to which the subject enjoys no particular epistemic authority. This is a theory of self-knowledge that assumes a 'self/other parity' (Schwitzgebel, 2014, Sect. 2.1).[4]

In social psychology Bem's self-perception theory pioneered a self/other parity account of self-knowledge. With reference to Skinner's methodological guidance, but with a position that reveals affinities with symbolic interactionism, he holds that 'individuals come to "know" their own attitudes, emotions, and other internal states partially by inferring them from observations of their own overt behavior and/or the circumstances in which this behavior occurs' (Bem, 1972, p. 5). This is reminiscent of Gilbert Ryle's well-known passage:

> The sort of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same […] in principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe. (Ryle, 2009, p. 139)

Nisbett and Wilson developed Bem's approach, claiming that behavioral and contextual data are the input of mechanisms that exploit theories that apply to the same extent to ourselves and to others. In one application of the so-called 'actor-observer' paradigm, Nisbett and Bellows (1977) compared the introspective reports of participants (actors) to the reports of a control group of observers who were given a general description of the situation and asked to predict how the actors would react. Observers'

[4] See also Robbins (2006, p. 619), who calls this account an 'outside access view of introspection'.

predictions were found to be statistically identical to—and as inaccurate as—the reports by the actors. This finding suggests that 'both groups produced these reports via the same route, namely by applying or generating similar causal theories' (Nisbett & Wilson, 1977, pp. 250–251; see also Schwitzgebel, 2014, Sects. 2.1.2 and 4.2.1). In this formulation, the self/other parity account of self-knowledge was welcomed by the proponents of 'theory-theory' in developmental psychology (see Gopnik, 1993).[5]

However, the self/other parity account is never suggested as an *exhaustive* theory of self-knowledge. Integral skepticism about introspection can hardly be found: some scope is always left for some sort of *direct* self-knowledge (see Schwitzgebel, 2014, Sect. 2.1.3). Nisbett and Wilson, for instance, draw a sharp distinction between *process* and *content*, that is, between the causal processes underlying judgments, decisions, emotions, sensations and those judgments, decisions, emotions and sensations themselves. Subjects have direct access to this mental content, and this allows them to know it 'with near certainty' (Nisbett & Wilson, 1977, p. 255). By contrast, they have no access to the cognitive processes that cause behavior. However, insofar as these authors offer no hypothesis about this alleged direct self-knowledge, their theory is *incomplete*.

In order to offer an account of this supposedly direct self-knowledge, some philosophers have tried to develop an up-to-date version of the Lockean 'inner sense' theory, construing introspection as a process that permits access to at least some mental phenomena in a relatively direct and non-interpretative way (Carruthers, 2011, Chap. 7). On this perspective, introspective access does not appeal to theories that serve to interpret behavioral and contextual data, but rather exploits mechanisms that can receive information about inner life through a relatively direct channel.[6]

[5] According to theory-theory, our folk-psychological abilities depend on the deployment of a 'theory' of the mental realm. The concept of theory involved has been unpacked essentially in two ways: (1) a body of knowledge which is stored in one or more innate modules, and gradually become functional ('mature') during infant development; (2) a body of knowledge that has much the same structure as a scientific theory, and it is acquired, stored, and used in much the same way that scientific theories are.

[6] This perspective is also called the 'inside access' view of introspection in Robbins (2006), p. 618; and the 'self-detection' account of self-knowledge in Schwitzgebel (2014), Sect. 2.2.

The attempt to bestow psychological plausibility on the inner sense theory of introspection comes in various forms. Introspection may be realized by a mechanism that processes information about the functional profile of mental states, their representational content, or both kinds of information (Robbins, 2006, pp. 618–619). Nichols and Stich's (2003) account of introspection in terms of monitoring mechanisms is a representationalist-functionalist version of the inner sense theory. Their hypothesis is that whereas detecting others' mental states and reasoning about one's own and others' mental states are all subserved by the same 'theory of mind information', the mechanism for detecting one's own mental states is quite independent of the mechanism that deals with the mental states of other people. More precisely, Nichols and Stich's hypothesis assumes the existence of a set of distinct self-monitoring computational mechanisms, including one for monitoring and providing self-knowledge of one's own perceptual states, and one for monitoring and providing self-knowledge of one's own propositional attitudes (henceforth 'thoughts').

The monitoring mechanisms account is concerned only with mentalistic self-attribution. As for third-person mentalization and third- and first-person mentalizing reasoning, Nichols and Stich make appeal to the theory-theory, allowing them to restrict the scope of experiments that show confabulation effects. The errors made by the participants do not concern mental-state self-attribution but rather first-person mentalistic reasoning; that is, understanding the causes of one's own behavior involves reasoning about mental states, and this is definitely a theory-laden process. Thus, if folk-psychological theory lacks the resources to account for a behavioral sequence, the participant will make inferential errors regarding both her own inner life and that of others. In other terms, self-knowledge can count on two methods: in some circumstances, individuals interpret by exploiting a folk-psychological theory which may give rise to confabulatory discourse; in other circumstances they can directly and non-interpretatively access their own minds.

Nichols and Stich see introspection as an inner sense faculty, that is, a faculty that provides us with a direct quasi-perceptual channel of informational access to our own mental life. This is also the view of Alvin Goldman's (2006), who, however, tries to relaunch the idea of inner sense within the framework of mental simulation. Here introspection both ontogenetically precedes and grounds mindreading. Mindreaders need

to introspectively access the offline products of their mental simulation before they can project them onto the target, and this is a form of direct access. Building on Craig's account of interoception, as well as Marr's and Biederman's computational models of visual object recognition, Goldman maintains that introspection is a perception-like process that involves a transduction mechanism that takes neural properties of mental states as input, and outputs representations in a code that represents types of mental categories. In the same vein, the hypothesis has been put forward that we mentally induce the internal states of others in ourselves through neuronal resonance (e.g., Gallese, Keysers, & Rizzolatti, 2004).

We have thus far discussed the approach of various philosophers who acknowledge the theoretical, and hence non-introspective, character of first-person knowledge of the causes of our thoughts and behavior, and nevertheless continue to think that in some specific cases the access to one's mental life is direct and non-interpretative. However, as we will now see, inner-sense theories are vulnerable to Peter Carruthers' criticism of the idea of a non-interpretative access to thoughts.

4.1.3 Self-Interpretation Plus Sensory Access

In opposition to the attempt to develop a cognitively plausible inner sense view of introspection (both in Nichols and Stich's version as well as in Goldman's), Peter Carruthers (2011, 2015) has developed a very sophisticated version of the self/other parity account: the Interpretive Sensory-Access (ISA) theory of the nature and sources of self-knowledge.

According to the ISA theory, we can have non-interpretive access only to a very limited range of sensorily accessible states; all knowledge of our own *occurrent thoughts*[7] is instead a matter of interpretation. More precisely, in agreement with the self/other parity account, our knowledge of our own thoughts is always the outcome of a swift and unconscious process of self-interpretation that exploits the same sources of evidence that we utilize when working out the mental states of others.

[7] That is, propositional attitude events (such as 'judging something to be the case', 'deciding to do something', or 'actively intending to do something') which are (1) *episodic* rather than persisting, and (2) have a non-sensory format (*amodal*) (see Carruthers, 2017, Sect. 1). Note, however, that there can be *sensorily-embedded* judgments, that is, judgments such as 'I see a reddish cat' which are directly grounded in sense perception (see Carruthers, 2015, Chap. 3, Sect. 5).

In order to account for the conscious accessibility of our perceptual states, the ISA theory assumes the validity of the above-mentioned global workspace models of human neurocognitive architecture. As described in Sect. 3.1.2, when one of the modular subsystems accesses the global neuronal workspace, its outputs (i.e., sensory information including perceptions of the world, the deliverances of somatosensory systems, imagery and inner-speech) are broadcast to an array of executive, conceptual and affective systems. These systems 'consume' sensory information to draw inferences, form memories, generate affective reactions, form judgments, plan and make decisions and enable verbal reports.

Among the conceptual systems that form judgments (which are largely events of belief formation) there is a single mindreading faculty, which is composed of a number of distinct but interacting parts. It exploits a corpus of folk-psychological theoretical knowledge in order to generate metarepresentational beliefs about the mental states of others and of oneself. This faculty has access to all sensory information broadcast by our perceptual systems, and hence has non-interpretive ('recognitional') access to one's own sensory states.

By contrast, thoughts—which are the outputs of the conceptual systems arranged in parallel around the global broadcast of attended perceptual information—are not capable of being globally broadcast, and hence, can never be consciously accessible. The reason is that global broadcasting depends upon top-down attention directed at mid-level *sensory* processing areas of the brain, implying that only mental events with a sensory-based format are capable of becoming first-order access-conscious (see Carruthers, 2015, Chap. 3).

In addition, top-down attention directed at mid-level perceptual regions of the brain is necessary not only for conscious perception but also in order that contents may enter *working memory* (thereby becoming access-conscious) (see Carruthers, 2015, Chap. 4). Since working memory is the system that underlies conscious forms of reasoning and decision making, all conscious *reflective* processes—as opposed to unconscious *intuitive* processes—must be sensory based.

As there are good reasons for thinking that there are no causal pathways from the outputs of the consumer systems to the mindreading system (see Carruthers, 2011, Chap. 3, Sect. 1.3), the latter must exploit the globally broadcast perceptual information, together with some forms of stored

knowledge, to infer the mindreader's thoughts, precisely as happens in the reading of other minds. Thus, as anticipated, the self-attribution of thoughts always occurs by means of a process of *self-interpretation*, which rests on the sensory awareness of data concerning one's own behavior, contextual data and/or sensory items in working memory.

In addition to experimental findings about the nature and sensory basis of broadcasting and working memory, Carruthers defends his ISA theory by taking a position on two areas where the ISA and inner sense theories make very different predictions.

The first area concerns the alleged dissociations between self-knowledge and other-knowledge in autism and schizophrenia. Since Nichols and Stich's (2003) monitoring mechanisms account assumes that introspection does not involve mechanisms of the sort that figure in mindreading, it implies that the first capacity should be dissociable from the second. Accordingly, they make the hypothesis of a double dissociation between schizophrenia and autism. In adults with Asperger's syndrome, the capacity of detecting their own mental states would be intact despite the mindreading deficit; the opposite pattern would be observed in schizophrenic patients with passivity experiences.

The ISA theory predicts that this dissociation should not occur, since there is just a single faculty involved in both mindreading and introspection. Consequently, Carruthers (2011, 2013b, Chap. 10) recruits data that refute Nichols and Stich's hypothesis. For example, two experiments conducted by Williams and Happé (2010) show that in children with Autism Spectrum Disorder (ASD) the capacity to attribute intentions to themselves is just as impaired as is the capacity to attribute intentions to others, and that the poverty of both performances can be imputed to the difficulties that ASD children have with mindreading in general. With regard to schizophrenia, Carruthers rightly points out that there is now extensive evidence of mindreading deficits in schizophrenia generally (see Brüne, 2005; Sprong, Schothorst, Vos, Hox, & van Engeland, 2007).[8]

[8] Carruthers also argues that passivity experiences can be much better explained by the hypothesis of the impairment of the so-called 'comparator system' (one of the main components of the action-control system) than by that of a system subserving first-person mentalization (see Frith, 2012; Frith, Blakemore, & Wolpert, 2000). However, there are experiments, such as the 'helping hands'

Even more than dissociation data, the ISA theory prediction which best serves to distinguish it empirically from inner-sense theories is that regarding frequent and pervasive confabulation, relative to one's own current or very recent thoughts.

Let us consider a classic case of confabulation for intentions (discussed in Carruthers, 2011, pp. 339–342). In Wegner and Wheatley's (1999) study, two subjects (a participant and the other a confederate of the experimenter) were invited to place their fingertips on a small board fixed to a computer mouse so that they could move a cursor around a computer screen that showed about 50 small objects from the book *I Spy*. Subjects were instructed to stop moving the mouse every 30 seconds or so, and then to rate on a scale from 0 ('I allowed the stop to happen') to 100 ('I intended to make the stop') the degree to which they had intended the stopping place of the cursor. Wearing headphones, the participant heard words (e.g., 'swan') that served to prime ideas about items on the screen. The confederate instead heard instructions on some of the trials to place the cursor on a certain object ('forced stops'), but was supposed to let the subject decide where to place the cursor on all other trials ('free stops'). For the forced stops, the subject heard the name of the target object via headphones either 30, 5 or 1 second before or 1 second after the stop.[9] Wegner and Wheatley found that participants rated their personal intention to stop on an object as higher when they had heard the name of the object 5 seconds or 1 second before they were forced to stop on it than when they heard the name 30 seconds before or 1 second after the forced stop. According to Carruthers, these results are just what the ISA theory would predict: when the word was heard just before the stop, the subjects' 'mindreading systems interpreted the coincidence of stopping near the object that had just been named as evidence of an intention to stop at that point' (2013b, p. 469).

---

study by Wegner, Sparrow, and Winerman (2004), which deeply challenge such a hypothesis by suggesting that a weak sense of agency can be elicited even when no motor prediction is formed.

[9] On Wegner's (2002) theory of apparent mental causation, if an action is consistent with a prior thought of the agent and other potential causes of the action are not present or salient, a sense of agency for the action is experienced. However, the thoughts must appear within a particular window of time for such an experience to develop. The variable manipulated through the I-Spy experiment was such a window of time—viz. the temporal interval between hearing the word and the time of the stop.

As already stated, the inner-sense theorists try to accommodate this kind of confabulation data by postulating two methods: not only an introspective, but also an interpretive route to our own attitudes. Consequently, inner-sense theories are less simple than the ISA theory. But above all, unlike the dual-method hypothesis, the ISA theory can explain the *overall patterning* of the confabulation data (see Carruthers, 2011, pp. 6 and 365–366). If the self-knowledge of thoughts is not direct, but results rather from self-directed mindreading, then there should be distinctive patterns of error in our claims about our thoughts, *mirroring the ways in which we can be misled about the thoughts of others*—for example, because the theory-driven interpretive process is fed by misleading sensory and behavioral data, or the theories that we use to interpret ourselves are inadequate.[10] The dual-method theorists, instead, will have difficulty in providing any principled account of the circumstances in which people access their thoughts directly, or of the circumstances in which they rely on self-directed mindreading.

4.1.4 Remnants of Introspection

Now we are in a condition to take stock of what we know on the nature of introspective self-consciousness.

We began with a large amount of data from neuropsychology and social psychology suggesting that the causes of our behavior and thoughts are not the sorts of things to which we have introspective access: causation cannot be introspected. If a person carries out a simple action, she will not able to report the complex chain of underlying motivational factors. Nevertheless, she will find no difficulty in providing reasonable and socially acceptable motives to justify her actions. Explaining why we are doing something, or why we decided to do it, has nothing to do with introspection, at least insofar as the latter refers to a direct access to the causes of our behavior and thoughts. Rather, it is the capacity to offer ex post facto rationalizations of our behavior in response to demands for reasons. In this sense, the self/other parity research tradition can be said to force a revision of the classical, 'intellectualist' conception of conscious agent.

[10] For example, in the case of the I-Spy study subjects are 'influenced by the presence of outcome-related sensory cues occurring shortly before the outcome itself' (Carruthers, 2011, p. 341).

We then discussed the approach of those philosophers who acknowledge the theoretical, and hence non-introspective, character of first-person knowledge of the causes of our thoughts and behavior, and nevertheless continue to think that, in some specific cases, access to one's mental life is direct and non-interpretative—an approach that often leads to some variant of Locke's inner sense theory. Nichols and Stich's theory of introspection postulates mechanisms that are fed, through a relatively direct channel, by information about perceptual states and thoughts. Goldman argues that the mindreader needs to introspectively access the offline products of mental simulation before it can project them onto the target—and introspection is a perception-like process. These theories, however, are vulnerable to Carruthers' criticism of the idea of a non-interpretative access to thoughts.

Carruthers' main argument takes the form of an inference to the best explanation. He contends that a wealth of cognitive scientific evidence speaks strongly against introspective access to one's own thoughts and is most adequately explained by his ISA theory, according to which self-knowledge of thoughts involves turning the mindreading faculty toward oneself. He holds that the only difference between self- and other-knowledge of thoughts is that in one's own case, the mindreading faculty has more available information upon which to base its interpretation. In addition to using overt behavior, in one's own case it can also draw on a subject's affective, sensory, and quasi-sensory states such as visual imagery or 'inner speech' tokens that are globally broadcast in the mind-brain.

Note that within the ISA framework, consciousness plays 'a crucial coordinating function in the minds of humans and most other animals' (Carruthers, 2016). Indeed, it is only when information becomes globally broadcast/access-conscious that it is made available to a wide range of executive, conceptual and affective systems, and this 'enables all these systems (and thereby the organism as a whole) to become coordinated around a common focus' (ibid.). Nevertheless, the distinctive feature of the global-broadcasting mechanism is that it is sensory based. Consequently, outside of the broadly sensory domain (sensation, perception and affect) none of our mental states is ever conscious. In particular, there are no such things as conscious (non-perceptual) judgments,[11] no such things as conscious intentions, and no such things as conscious decisions.

[11] The qualification in brackets is required because, as said in note 7 above, there can be conscious *perceptual* judgments.

It is important to notice that this disappearance of conscious thought still leaves room for a distinction between *unconscious intuitive* processes and *conscious reflective* processes. The latter are forms of mental activity that are directed, for example, toward solving a problem, arriving at a judgment, or at reaching a decision. Since working memory is the system that underlies these reflective processes, our conscious reflections will be exclusively composed of sensory-like events.

4.2 The Construction of the Virtual Inner Space of the Mind

Within the ISA framework, our inner life consists in the unfolding of a lush perceptive phenomenology which relentlessly feeds a machinery of interpretation driven by an incomplete, partial and, in many cases, seriously defective naïve theory of psychology. Self-consciousness as introspective reflexivity is thus largely an activity of reappropriation of the outputs of the unconscious cognitive processing. Equipped with this result, we turn our attention to the ontogenesis of the inner, virtual 'theater' of the mind.

4.2.1 Mindreading and Attachment

In Sect. 4.1.4, we examined two areas, dissociation and confabulation data, where the ISA and inner sense theories of self-knowledge make very different predictions. Here, we consider a further area, which concerns the nature and source of our capacities for metacognitive control of learning and reasoning.

The ISA theory posits a single phylogenetic route for both mindreading and introspection—an integrated faculty of metarepresentation that evolved for mindreading and was later exapted for introspection. This is what is legitimate to expect in light of the hypothesis that mindreading, as an ingredient essential to our social intelligence, evolved to provide an adaptive advantage in pursuing the aims of the two motivational macrosystems discussed in Sect. 2.4, the first committed to self-assertiveness and

competition, and the second aimed to pro-sociality and cooperation. Our competence to mind-read others is then a cognitive adaptation designated to efficiently predict, interpret and manipulate the actions of other conspecifics in a variety of competitive as well as cooperative situations.

Buckner, Shriver, Crowley, and Allen (2009) made the objection that metarepresentational mindreading is likely to be a late exaptation of more primitive capacities (e.g., first-order, non-metarepresentational mechanisms for face recognition, eye-tracking, automated imitation via the mirror neuron system, and so forth), grounded in these together with our linguistic abilities and general-purpose concept-learning and theorizing skills. According to Carruthers (2009b), however, this view can be rejected based on two sources of evidence. First, after Onishi and Baillargeon's (2005) groundbreaking paper, enough evidence has accumulated to make plausible the hypothesis that a core form of metarepresentational mindreading is not 'a developmental achievement, but an innate social-cognitive evolutionary adaptation' (Gergely & Unoka, 2008a, p. 58; see also Cosmides & Tooby, 2013). Such adaptation is implemented by a neurocomputational system that begins to operate very early in life (Baillargeon, Scott, & Bian, 2016; Carruthers, 2013a). Second, metarepresentation is required for lexical acquisition; if children were not able to grasp a speaker's referential intentions, learning the meanings of words would not be possible (see Bloom, 2000).

In contrast with the ISA theory, a 'first-person based' account of the evolution of metarepresentational capacities suggests that the capacity to represent one's own mental states (or some subset thereof) was the first to appear in evolution, presumably to enable our ancestors to increase the advantages of metacognitive monitoring and control (see Couchman, Coutinho, Beran, & Smith, 2009). Once evolved, these first-person monitoring-and-control abilities were somehow exapted for mindreading.[12]

[12] This could have happened in one of two ways: 'Either these first-person resources were redeployed to form the basis of a distinct mentalization faculty of the sort defended by Nichols and Stich (2003), or they were combined with emerging capacities for imaginative perspective-taking to enable *simulations* of the mental lives of others' as Goldman (2006) suggests (Carruthers & Ritchie, 2012, pp. 78–79).

However, the hypothesis that introspection evolved for metacognitive purposes does not tally with the available evidence. The human and comparative metacognitive data seem to show at least two things. First, in many cases the controlling function of metacognition does not involve any self-directed metarepresentational capacity.[13] Second, even where our metacognitive interventions are metarepresentational, deploying concepts of mental state types (as in the cases of meta-reasoning, meta-learning and meta-memory), they seem to be incapable of the sort of direct impact on cognitive processing that would be predicted if metacognition had, indeed, evolved for that purpose (see Carruthers, 2009b, 2011, Chap. 9; Carruthers, Fletcher, & Ritchie, 2012; Carruthers & Ritchie, 2012).

* * *

The 'mindreading first' hypothesis is also fully consonant with attachment theory, since it is developed within a contextualist and systemic framework in which (individual) biology and (social) relationality cannot be separated (see Sect. 2.4). Individuals are pre-wired to the interpersonal relationship from birth, and the above mentioned early-developing core mindreading system is part and parcel of such pre-organization.

It is to be noticed, however, that the mindreading system is an innate social-cognitive adaptation that is independent of Bowlby's innate infant-caregiver attachment system. In this perspective, attachment and mindreading are two independent adaptations that have been selected to serve qualitatively different evolutionary functions. This is in contrast with the hypothesis, variously put forward by a number of attachment theorists and infant researchers, that there is a direct causal and functional link between secure attachments during the infant's first year on the one hand, and the development of mindreading on the other (see Gergely & Unoka, 2008a).

According to Meins (2011), however, the observed link between attachment security and mentalization may be *indirect*, with both attachment security and mentalization performance being predicted by caregivers' *mind-mindedness*, that is, the proclivity to treat one's infant as an individual with a mind, rather than merely an entity with needs that

[13] Even though the standard view is that metacognition involves self-directed metarepresentation (Nelson and Narens, 1990), a large number of metacognitive processes are not metarepresentational. One example is the above-mentioned (note 8) 'comparator system' which does not involve any metarepresentations.

must be satisfied. One aspect in particular of the caregivers' internal-state language, namely comments that appear to be appropriate to the mental state of the child, foretells children's future mentalizing performance.

Now, there is no doubt that caregiver-infant communicative interaction impacts on the development of mentalization; the problem is how the child's exposure to such interaction can have such an impact. In particular, the role that language plays in this context needs to be clarified. Jill and Peter de Villiers, for example, would disagree with Meins' hypothesis that language, in the form of comments that appear to be appropriate to the mental state of the child, is crucial as an element that is able to impact on the development of mentalization. More radically, they think that our metarepresentational mentalistic abilities are *constituted* by language; more specifically, the claim is that the mastery of the grammatical rules for embedding tensed complement clauses under verbs of speech or cognition provides children with a necessary representational format for dealing with false beliefs (for references, see de Villiers, 2013). However, such claim seems to be at odds with the evidence (see Carruthers, 2011, p. 226). For example, Perner, Zauner, and Sprung (2005) have shown that mastery of sentential complements is not a necessary condition of the development of mindreading in children. For such a mastery may be required for statements about beliefs but not about desires (as in English), for beliefs and desires (as in German), or for neither beliefs nor desires (Chinese); yet, children who learn each of these three languages all understand and talk about desire significantly earlier than belief. But above all, any theorizing on the relation between language and mindreading must come to grips with the already mentioned evidence suggesting that infants between the ages of 6 and 18 months are capable of representing and reasoning about the false beliefs of other agents. Such evidence knocks out a constitution-thesis *à la* de Villiers, but also raises a problem for Meins' (2011) view of the relation between language and mentalization: the attachment environment is a form of scaffolding that begins with proto-conversational exchange, and only later becomes linguistic.

Thus, insofar as mindreading is concerned, there is no direct ontogenetic causal and functional link between the quality of early infant attachment—or the linguistic scaffolding consisting in mothers' internal-state talk that is appropriately attuned to the infant's thoughts and feelings—on the one hand, and the development of mindreading on the other.

When we take introspection into consideration, in contrast, the relationship between attachment and mentalization is no longer simply one of scaffolding (see Marraffa, 2015; Marraffa & Meini, 2016). As we will see in the next section, the child's socio-communicative interaction with caregivers is *constitutively* involved in the construction of the inner experiential space.

4.2.2 The Construction of Introspection in the Attachment Environment

The importance of the attachment theory for the development of the child's introspective abilities is nicely highlighted in the social biofeedback theory of parental affect mirroring (see Fonagy, Gergely, Jurist, & Target, 2002; Fonagy, Gergely, & Target, 2007; Gergely, 2004; Gergely & Unoka, 2008a, 2008b; Gergely & Watson, 1996, 1999), a socio-constructivist model of the development of the virtual inner space of the mind. Within the attachment theory framework, a mother and child create a system of affective communication from the beginning of life, one in which interactions with the caregiver play a fundamental role in the modulation of the infant's affective condition. The social biofeedback theory holds that the caregivers' attuned and marked affective 'mirroring' in repetitive episodes of nonverbal communication is the beginning of a developmental pathway that, starting from an ebb and flow of core affects, leads to the construction of discrete emotions (i.e., emotion episodes designated by an individually separate and distinct category such as fear or anger — see Scarantino, 2014), and to their subsequent internalization into one's own inner life. Thereafter, we will examine how the construction of introspective self-knowledge makes headway through the linguistic scaffolding consisting in caregiver's internal-state talk that is appropriately attuned to the infant's thoughts and feelings.

* * *

The most common context of dyadic, affective relationships involving a child and her caregiver are turn-taking 'protoconversational' interactions: 'Both partners actively interact, reciprocally exchanging information during a conversation made up of imitations (but also of subtle

episodes of desynchronization), improvisations, search for eye-to-eye contact, sensitivity to vital forms, and so forth' (Meini, 2015, p. 285). One of the most advanced models of these interactive exchanges is the just mentioned social biofeedback theory of parental affect-mirroring.

The model is completely at odds with the 'strong intersubjectivist view', namely, the claim that infants are born with a pre-wired organization of their minds that ensures a primary introspective access to their own affective and intentional mental states (see Gergely, 2002). In Sect. 3.3, we cited an example of such a position, namely, Meltzoff and collaborators' hypothesis of a specific innate mechanism underlying intersubjective attributions during early imitative interactions. The affective behavioral acts of the other are mapped onto the infant's supramodal body scheme, allowing her to recognize the other person as 'just-like-me'. By imitating such acts, infants generate the corresponding feeling states in themselves; these are then *introspectively accessed* and attributed to the other by inference. This is in line with Goldman's inner-sense view that introspection both ontogenetically precedes and grounds third-person mentalization (see Sect. 4.1.3).

By contrast, the social biofeedback model makes the hypothesis that at the beginning of life human infants show a primary bias to attend to and explore the *external* reality, and construct representations mainly based on exteroceptive stimulation. Initially, therefore, the set of visceral and proprioceptive cues that are activated when being in and expressing an emotion state are 'not grouped together categorically in such a manner that they could be perceptually accessed as a distinctive emotion state' (Gergely & Watson, 1999, p. 110; see also Gergely & Unoka, 2008a, p. 62).

In short, there is no phenomenology of discrete emotions. In accordance with the 'differentiation' theories of early emotional development (e.g., Sroufe, 1996), such emotions are seen as emerging from simpler precursors. We are dealing here with an asymmetry between third and first person. On the one hand, early caregiver-infant interactions 'are characterized by frequent exchanges of a relatively rich and differentiated (and ontogenetically quickly increasing) repertoire of facial-vocal emotion displays expressing specific basic emotions (including anger, joy, fear, sadness, disgust and interest)' (Gergely & Unoka, 2008a, p. 53). On the other hand, in their initial state infants have only undifferentiated

internal experiences of positive and negative arousal; states of introspective awareness of specific basic emotions[14] are differentiated through interactions with the attachment environment.

Thus, in the first few months of life, infants are complex representational systems who are able to respond to the caregiver's display of various basic emotions.[15] In the Darwin-Tomkins-Ekman tradition, basic emotions (the most elemental among discrete emotions) are biologically based and pancultural packages of short-term, coordinated and automated responses to events in the environment, which include a somatic component (e.g., measurable physiological changes), a motor component (e.g., facial and vocal expressions) and a motivational component (i.e., action tendencies). These responses are assumed to be automatically elicited and coordinated by a causal mechanism called the 'affect program' (e.g., Ekman, 1999; Panksepp, 1998).

Yet, at this stage infants lack the *feeling* component of such discrete emotional states; there is no reflection of (aspects of) the other components of basic emotions into first-order consciousness. As argued in Sect. 3.3, in its initial state the human organism's experiential space is purely objectual, and the original form of differentiation of this buzzing and blooming space is likely to occur in accordance with a basic alternation of our dispositional orientation toward reality, that between the *positive* and *negative* affects.

Here comes into play the notion of valence. Valence is an elemental, binary, antinomic dimension of the agent's dispositional orientation toward reality. It combines an appraisal component (bad/good, intended as unpleasant/pleasant) with an arousal component (more or less excited). Thus, in assessing any attitude toward an object, or a class of objects (attitudes of rejection, acceptation, suspicion, aggression, love, jealousy and so on), a valence (or 'direction', or 'sign') can be assigned to that attitude, locating it at a point on a line that has, at one end, a total and enthusiastic affective involvement in the object, and at the other a total and aggressive rejection of it, with indifference at its center (see Russell, 1980, 2003).

[14] For example, 'awareness of being 'angry', rather than just experiencing some undifferentiated negative state of tension' (Gergely, 2004, p. 58).

[15] Of course, the subject matter of emotions is huge and there are several different theories of the nature and ontogeny of emotions on the market. However, we need not discuss them here; for our purposes it is sufficient to introduce the notions of basic emotions and valence.

In all animals the relationship with the external world is mediated by a basic alternative that is of this kind, namely, 'approaching/withdrawing', 'accepting/rejecting', 'incorporating/expelling'. Even the simplest animals deal with objects and events according to the 'good/bad' dyad. The same holds for newborn infants. From birth, primary appraisals of the world enable the infant to discriminate whether an object or situation is helpful or harmful, rewarding or threatening, requiring approach or withdrawal. And the presence of positive valence will give rise to feelings of acceptation-pleasure-reassurance-incorporation, whereas the presence of negative valence will give rise to feelings of rejection-insufficiency-distress-expulsion. It is on the basis of this kind of fundamental distinction that the newborn infant begins to organize a relationship with the world.

* * *

Within this framework, the social biofeedback model of parental affect-mirroring provides an account of how infants move from an undifferentiated affective state, characterized merely by valence, to an awareness of one's own discrete emotional states.

Gergely and Watson (1996, 1999) posit an innate contingency perception mechanism that enables the infant to analyze the *conditional probability* of three contingent relations—temporal contingency, spatial similarity and correspondence of relative intensity—between own actions and effects in the external environment.[16] The social biofeedback model applies this hypothesis to the special case of parental affect-mirroring.

At birth, infants are unable to regulate their own emotions. The infant's acquisition of an ability for emotional self-control depends on a species-specific characteristic of the human attachment system: the inclination of sensitive, infant-attuned caregivers to provide an emotional scaffolding environment by mirroring back the infant's affect-expressive displays in a 'marked' way. This means that the facial-vocal pattern of such displays is a schematic and exaggerated version of the corresponding realistic emotion expression of the caregiver.

[16] The mechanism uses two different and independent indices for estimating the degree of causal relatedness between responses and stimuli. One of these indices is the 'sufficiency index' which registers the conditional probability that a certain stimulus (A) will be followed by a certain response (B). The other, the 'necessity index', monitors the likelihood that a given response B was preceded by a stimulus A (see Gergely & Watson, 1996, pp. 1190–1196).

The facial and vocal exaggeration of the parental mirroring, coupled with her soothing tone, serves to mitigate the potentially arousing effect of direct imitation, while simultaneously making salient to the infant central aspects of the somatic manifestations of an emotion (see Kim, Fonagy, Allen, Martinez, Iyengar, & Strathearnet, 2014). The markedness of the parental affect-mirroring display signals to the infant that the displayed emotion is 'not for real', and that its dispositional content should be referentially 'decoupled' from the caregiver. This interpretation is supported by two other dimensions of the affect-mirroring displays: (1) the suspension of negative behavioral consequences for the infant when faced with a mirroring of negative emotions by the caregiver; and (2) the high degree of contingent relatedness between the infant's emotion expressive facial-vocal responses and the parental affect-mirroring expressions which is detected by the contingency perception module.

Thus three features—markedness, nonconsequentiality and high contingent relatedness—differentiate 'as-if' (or pretend) emotion communications from realistic emotion displays. As a result, the infant will set up *separate representations* for the affect-mirroring displays. This leads to the above mentioned referential decoupling of the affect-mirroring display from the caregiver—in other words, it will be represented as 'not being about' the caregiver's actual emotion state. Once decoupled, however, the affect-mirroring display still needs to be interpreted by the infant as referring to 'someone's emotion'. This process of 'referential anchoring' depends on the infant's contingency-detection system which registers the high degree of contingent relation between the parental mirroring and the infant's ongoing affective behavior. As a result, 'the infant will *referentially anchor* the marked mirroring stimulus as expressing his *own* self-state' (Gergely & Watson, 1996, p. 1199).

Thus the parental affect-mirroring serves mainly two functions. A function of *sensitization*: the infant becomes sensitive to the set of internal physiological and proprioceptive cues that are active while her affect-expressive behavior is controlling the adult's marked affect-mirroring expressions. A function of *representation building*: the separate representations of the caregiver's affect-mirroring displays become associated with the infant's primary and procedural affective states; thus they form *secondary representations* that are about those primary affective states and provide the basis for the infant's emerging ability to control her emotion states (see Gergely, Koós, & Watson, 2010, Sect. 2.5).

Drawing it all together. The earliest form of differentiation of the infant's experiential space occurs in virtue of valence dimensions: positive valence will give rise to feelings of acceptation-pleasure-reassurance-incorporation, whereas negative valence will give rise to feelings of rejection-insufficiency-distress-expulsion. In contrast, there is no phenomenology associated to basic emotions. In the initial stage, basic emotions are packages of somatic, motor and motivational components elicited and coordinated by causal mechanisms (affect programs) which play the role of social signals in the 'negotiation' between infant and caregiver (see Griffiths & Scarantino, 2009, p. 446). It is affect mirroring that adds a phenomenological component to basic emotion packages. As seen, marked mirroring displays are interpreted self-referentially by the infant, leading to their referential anchoring (in the form of internalized second-order representations) to those procedural basic emotion states that the mirroring displays contingently reflect. This process will lead to the internalization of discrete emotions into the infant's own inner life when—in the second year of life—the phenomenology of basic emotions is embedded into bodily self-consciousness, making the infant's bodily self-image an *affective* bodily self-image (see Marraffa & Meini, 2016, Sect. 3.1).

Two aspects of this socio-constructivist approach to affective introspection are particularly important for our purposes. First, knowledge of one's own mind rests on interactions with the attachment figure, who displays emotional expressions of which the child already knows the meaning. Already at this basic level, therefore, we find the primacy of third-person cognition: at any level of complexity, knowledge of the self requires at least an equivalent level of knowledge of others.

Second, this approach to the emergence of the awareness of one's emotions supports our claim that bodily self-consciousness is a necessary premise of the further development of the ability to identify the presence of an inner experiential space. Although we are not committed to the Jamesian idea that all emotions are perceptions of aroused states of the body (Damasio, 1999; Prinz, 2004), the earliest cognition of mental events appears to be the outcome of the acquired capacity of 'interpreting primary somatic data specific to categories of affective states and of attributing them to the self' (Hernik, Fearon, & Fonagy, 2009, p. 148). In this view, emo-

tions are events that are originally detected in the body, and subsequently internalized in our mental life. Section 4.3.1 shows that it is this internalization that enables us to responsibly take possession of emotion episodes.

* * *

The social biofeedback model is a contribution to a theory of mentalization that aims to go beyond the classical construct of theory of mind with its associated false beliefs paradigm. Such a construct is too narrow, 'as it fails to encapsulate the relational and affect regulative aspects of interpreting behavior in mental state terms' (Fonagy, Gergely, & Target 2007, p. 288). In this perspective, an integration is required between Carruthers' view of the introspection of thoughts as self-directed mindreading, and Gergely and Watson's socio-constructivist approach to the developmental mechanisms by which mentalizing abilities give shape to the virtual inner space of the mind.

Although Carruthers makes a strong case for the claim that mindreading has a functional and evolutionary priority over introspection, his theory of introspective self-knowledge does not predict that mindreading should also be *developmentally* prior to introspection (2009b, p. 167). However, in an attempt to explain why we have the (false) intuition that there is introspection for our thoughts, the philosopher takes very seriously Wilson's (2002) hypothesis that the self-transparency assumption 'may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image' (Carruthers, 2009a, p. 138, n. 5). Moreover, in examining the possibility that the emergence of introspection is a by-product of the evolution of mindreading, Carruthers considers such a possibility as compatible with the hypothesis that introspection 'might have come under secondary selection thereafter, perhaps by virtue of helping to build and maintain a positive self-image, as Wilson […] suggests' (ibid., p. 128).

Thus, a door is opened here to the topic of defense mechanisms, that is, the hypothesis that our activity of re-appropriation of the products of the unconscious is ruled by a self-apologetic defensiveness. And that is how it should be since the ISA theory draws heavily on the confabulation data from the huge cognitive dissonance and causal attribution literatures (Carruthers,

2011, Chap. 11), and such data can hardly be separated from the topic of the construction and maintenance of 'a positive self-image' (see Sect. 5.2.1).

There is a problem, however. Carruthers' focus is *not* on self-knowledge construed as 'awareness of oneself as an ongoing bearer of mental states and dispositions, who has both a past and a future' (Carruthers, Fletcher, & Ritchie, 2012, p. 14). His focus—as he makes clear—is on *knowledge of one's own current mental states*; and this knowledge 'is arguably more fundamental than knowledge of oneself as *a self with an ongoing mental life*' (ibid.; italics added). However, insofar as introspection is taken *merely* as a competence to self-attribute one's own current mental states, Wilson's hypothesis of the self-defensive nature of introspection cannot be built into the ISA theory. For, as will be made clear in Chap. 5, the topic of defenses makes sense only in the context of the construction and protection of the psychological self-consciousness or narrative identity ('a self with an ongoing mental life'). But once introspection is seen in this context it becomes possible to make the hypothesis that it *develops* through the act of turning upon oneself the competence to mindread others, and that this occurs through the socio-communicative interaction with caregivers (and subsequently, other social partners) investigated in the attachment theory research.

Two different types of introspection appear to be at stake in an exchange between Carruthers (2009b) and Fernyhough (2009). Fernyhough draws attention to some sources of evidence for the hypothesis of a late emergence of the child's inner experience—in particular, findings indicating that the transformation of 'private speech' (i.e., speech that is not obviously addressed to any interlocutor) into inner speech may not be complete until middle childhood, and that visual imagery also takes time to develop (e.g., Al-Namlah, Fernyhough, & Meins, 2006; Fernyhough, Bland, Meins, & Coltheart, 2007). Since inner speech and visual imagery are among the data that feed the interpretive process underlying knowledge of one's propositional attitude states, Fernyhough concludes that the emergence of introspection would have to be developmentally constrained by the emergence of inner speech and visual imagery. Given what we know about the timetable for the emergence of mentalizing

abilities (especially the already mentioned evidence for very early mindreading competences), the ISA theory should predict a developmental lag between mindreading and introspection. However, Carruthers has denied any such implication.[17]

The problem here seems to be that Carruthers and Fernyhough are approaching introspection from very different perspectives. As said, the former's focus is on a minimal sense of introspection as competence to self-attribute one's own current mental states taken independently from any cognition of oneself as a self construed as introspective self-description, that is, psychological self-consciousness or narrative identity. By contrast, Fernyhough's focus is on the development of introspective self-consciousness in a Vygotskian perspective: an outward-in construction that occurs in an interpersonal context, namely in the relationship with caregivers and peers. Thus, Carruthers takes introspection as a competence *in isolation*, and this notion is 'too restrictive' to elaborate our understanding of its development beyond 'the standard strategy of comparing children's performance across false-belief tasks' (Hernik, Fearon, & Fonagy, 2009, p. 147). Fernyhough, in contrast, analyzes introspection within a framework in which the turning of one's mindreading abilities upon oneself is seen as part of the construction of an inner experiential space, and then of a narrative identity. It is introspection taken in this constructive dimension that is relevant to the psychodynamic topic of defenses.

To recapitulate, we began by taking a nativist-modularist perspective on mindreading, endorsing the hypothesis that a form of primary mindreading is not a developmental achievement, but rather an innate social-cognitive evolutionary adaptation implemented by neurocomputational mechanisms that are active and functional by the first year of age. Subsequently, endorsing Gergely and Watson's criticism of the claim of primary intersubjectivity, we adopted a social-constructivist stance on introspection. Finally, expanding on Carruthers' strong case for the

[17] All that follows, Carruthers writes, 'is that there will be many more moments in the daily lives of children at which they will be unwilling to attribute occurrent thoughts to themselves than is true of the daily lives of adults, because the conscious mental events that might underlie such self-attributions simply are not present. Nothing follows about children's competence to self-attribute attitudes. Nor does it follow that children will be weaker at attributing attitudes to themselves than they are at attributing attitudes to others, provided that the tasks are suitably matched' (2009b, p. 167).

claim that mindreading has a functional and evolutionary priority over introspection, we maintained that mindreading is also developmentally prior to introspection. If this competence is placed in the context of 'the relational and affect regulative aspects' of mentalizing, good reasons emerge for arguing that one of the factors of its development is the act of turning upon oneself the capacity to mentalize, and that this occurs through the socio-communicative interaction with caregivers (and subsequently other social partners) investigated by attachment theory. As Gergely and Unoka suggest, 'with the developmental construction of cognitively accessible second-order representations of internal self states, the proper domain of the human mindreading becomes ontogenetically extended to include in its actual domain the mind of one's own self as well' (Gergely & Unoka, 2008a, p. 74).[18]

* * *

After internalizing basic emotions into one's own unfolding inner life, the child must learn to recognize and attribute to herself other kinds of mental states and activities, as well as forming the conceptual network that links such phenomena. In other terms, she must develop a genuine theory of one's own mind. This occurs in a context in which the socio-communicative interaction with caregivers moves from the preverbal to the verbal stage. As a result, a whole new range of mature mentalistic activities—which exploit the basic mentalistic abilities underpinned by the early-developing mindreading mechanism—emerges under the thrust of caregivers' mind-minded talk (see Meins, 2011; Nelson, 1989, 2007).

Here comes into play a component of the mindreading system that systematically reads other people's behaviors as actions driven by goals, purposes, intentions—intentions modulated according to the 'good intention vs. threatening intention' dichotomy (e.g., Hamlin, 2013a, 2013b). The question 'What does *that* want to do?'—where 'that' can refer to the mother or the home cat—is early and primary. And then, on the basis of this kind

[18] On the distinction between the 'proper' and 'actual' domains of an evolved cognitive system, see Sperber and Hirschfeld (2004). On the one hand, the specialized system evolved to represent and react to a set of objects, facts and properties; on the other hand, the system *actually* reacts to a set of objects, facts and properties. According to Gergely and Unoka, thus, the proper domain of mentalization was originally restricted to inferring and representing the causal intentional mental states of other minds only.

of problem, the child begins to ask *also* what her own intentions are, and what her own inner state is. This appropriation of themes that were initially connected only to the reading of others' behaviors is mediated mainly by a learning that is educational, and hence cultural. In other words, it can be supposed that most of the simplest introspections are forms of learning emerging from the verbal stereotypes and rhetoric through which adults rename the intentions of others. A two-year-old child, perhaps because she is frightened by her granny's cat, perhaps as an act of defiance, gives the cat a boot, and here follow the reconstructive judgments about this episode on the part of the adults, which she is invited to internalize: 'Bad child! It didn't mean to claw you at all!', or 'It scared you, but perhaps the kitty was more scared than you'. And so the young child gradually learns—always internalizing the (hypothetical) names that the adults give to her inner states—that inside her there are scares, badness, and so on. She understands that these are contingent social expressions, part of social mediations, but also grasps what 'information about herself' means.

Note again the connection between the construction of interiority and ethics, on which Locke had already drawn our attention. Morality reinvents interiority from scratch: being bad and being good, having bad intentions and having good intentions, appear to the child the premise of imputability even before responsibility. This permits the explanation of why, despite the above-mentioned verbal stereotypes and rhetoric actually containing a plea for responsibility, in our culture the sense of *responsible* appropriation of one's own actions—so that I know that I could be objectively and legally responsible for a car accident even if I am not able to identify in myself an intention to cause it—[19] is usually replaced by a less clear and more sterile feeling, the sense of guilt. For the sense of guilt can be ascribed precisely to that instinctive-primitive interpretation of human actions which always and necessarily links them to an aware intentionality, good or bad, and makes it difficult to accept and understand the presence of involuntary, fortuitous, inattentive or unaware behaviors. That an action can be an offence irrespective of good or bad intentions is not taken into account by our folk psychology.[20]

[19] In regards to this matter, legal language distinguishes between culpable unintentional antisocial acts, whose damaging effects can be ascribed to an agent who has not planned them as such, and malicious acts, where, on the contrary, there existed the (plan) intention to reach that outcome.

[20] We return to the nexus between interiority and ethics in the Epilogue.

Thus introspective self-consciousness takes shape in the child in the context of her relationship with the caregiver—a relationship that is made first of preverbal proto-communicative exchanges, and then of words, descriptions, designations, evaluations of the person. Through such interaction with caregivers (and then with other social partners) children construct their own identity, both *objective* (for others) and *subjective* (for themselves). And identity-for-oneself can be said to arise out of identity-for-others; introspective self-description takes shape through a creative process of internalizing the ways in which others see and define us.

This hypothesis made by Mead (1934) is still fundamental to an understanding of the relationship between individual differentiation and social belonging. As Gergely argues, placing himself within the tradition of social constructivism in self-development,

> …the intentional actions and attitudes repeatedly expressed towards the young child by caregivers and peers serve as the inferential basis for attributing generalized intentional properties to the self in an attempt to rationalize the social partners' self-directed behavior. This is how the establishment of a 'categorical' self-concept or representation (the Jamesian 'Me') originates. (Gergely, 2002, p. 42)

Here it is worth noting how psychology amended a merely sociological construal of Mead's hypothesis. Such a construal has had the drawback of underestimating the complexity, the fatigue, the creative aspects and the risks of the internalization process. By contrast, developmental, social and dynamic psychology have steered Mead's hypothesis onto the right path, making it clear that infants are *active creators* not only of their structures of relationship with other people, but also of their ways of self-presentation.

4.3 The Emergence of a Continuous Self Through Time

Subjective identity evolves. The child gradually comes to experience herself as a person, to define herself as a certain kind of person, and to trace her own continuous identity as a person across time and space

(see Fivush, 2010). As James puts it, subjective identity consists in finding oneself again among the intermittences of consciousness: 'Each of us when he awakens says, Here's the same old self again, just as he says, Here's the same old bed, the same old room, the same old world' (James, 1950 [1890], p. 334).

A sense of 'temporally extended self' (Povinelli, 1995) or 'self in time' (Nelson, 1989) is a complex cognitive achievement, which originates from the establishment of an autobiographical memory system. Children are required to develop the capacity to perceive their identity as situated in memory: they must be able to represent not only the 'what', 'where', and 'when' of a past event, but also themselves as the subjects who experienced that event. This perception of an identity situated in memory will be progressively rationalized in autobiographical terms.

One way of defining autobiographical memory is in terms of episodic memory; autobiographical memories are a special subclass of episodic memories, that is, those that involve a reference to the self or are of a particular kind of relevance to the self. In Tulving (2002) such a reference to the self is built into the very notion of episodic memory. Episodic memory is accompanied by *autonoetic* consciousness, which provides 'a recollective experience infused with a sense of one's self extended in time' (Prebble, Addis, & Tippett, 2013, p. 4). In contrast, *personal semantic* memory is a *noetic* form of memory, which is associated with a feeling of 'knowing' rather than remembering one's own life (see Addis & Tippett, 2008, p. 73). In this theoretical framework, although both types of autobiographical memory involve a reference to the self in the sense that their contents relate to 'my' past, it is episodic memory that is intimately connected to a sense of self, because it is supposed to entail 'a direct, intimate, and immediate sense that "I" experienced the event' (Prebble, Addis, & Tippett, 2013, p. 4; see also Vandekerckhove, 2009; Wheeler, Stuss, & Tulving, 1997; Zahavi, 2005). As Tulving puts it, 'episodic memory differs from other forms of memory in that its operations require a self. It is the self that engages in the mental activity that is referred to as mental time travel: there can be no travel without a traveler' (Tulving, 2005, pp. 14–15). From here, it is only a very small step to explain episodic memory in terms of the construct of pre-reflective self-consciousness which was challenged in Sect. 3.3.

The step is taken by Prebble, Addis, and Tippett (2013), who see it as 'an elegant solution' to the classic problem of self-continuity or 'diachronic unity'. It is pre-reflective self-consciousness ('the phenomenological flavor of mineness through time', as they say on p. 829) that is the precondition for episodic autobiographical memory, which in turn, because of its above-mentioned qualities of autonoetic awareness and mental time travel, is a prerequisite for experiencing unity in our subjective experience of selfhood across time—or 'phenomenological continuity' (see Sect. 3.4). Our ability to remember episodically solves the problem of diachronic unity insofar as it carries 'the inherent 'mineness' of the original experience into the present moment' (ibid., pp. 818–819). The next developmental step will be the gaining of a sense of 'narrative continuity', which depends mainly on semantic autobiographical memory (see also Addis & Tippett, 2008).

Once more, however, appealing to pre-reflective self-consciousness takes us in the wrong direction. In the first place, the prerequisite for autobiographical entry and storage seems to be bodily self-awareness.

Most of the theories of autobiographical memory development have been cast in terms of explaining infantile amnesia, the phenomenon by which adults cannot recall most of their early childhood experiences. According to Howe and Courage (Howe, 2011, 2014; Howe & Courage, 1993, 1997; Howe, Courage, & Rooksby, 2009), before the preschool period, children lack a critical cognitive or social-cognitive framework that would enable them to encode and store memories in such a way that they could later be retrieved as relevant to the self. This framework is *self-consciousness* as commonly measured in the mirror task of self-recognition. Awareness of self is thus responsible for 'kick-starting' autobiographical memory:

> This is because, now that the self has recognizable features (e.g., sensations, feelings), it can serve to organize and structure experiences in memory. Before this, experiences were simply remembered as events that happened, events that were only loosely bound in relatively fragmented trace structures. With the advent of self-consciousness, the events that are now being experienced become personalized, in the sense that they are now events that happened to this self, events that happened to 'me'. (Howe, 2014, p. 552)

Now, we agree with Howe and Courage that the most important factor in the emergence of autobiographical memory is self-consciousness as measured in the mirror self-recognition task. However, we take issue with the authors' construal of the fixed referent as a 'cognitive self-concept', because it assumes the strong mentalistic interpretation of mirror self-recognition which was discarded in Sect. 3.3. Our sense of ourselves in time is rooted in the onset of a *physical* form of self-describability: the nonverbal, analogic representation of the bodily self, constructed in the second year of life, acts as a fixed referent around which personally experienced event memories begin to be organized. In James' terms, the Me to which the infant begins to attach episodic memories is the material self.

In the second place, Prebble et al. (2013, p. 819) think that their framework (see Sect. 3.4) can include both their account of episodic memory in terms of pre-reflective self-consciousness and Martin Conway's model of the interconnectedness of self and memory (see Conway, 2005; Conway & Pleydell-Pearce, 2000; Conway, Singer, & Tagini, 2004). But this cannot be the case; let us see why.

In Conway's cognitive-motivational model, autobiographical memories are generated within a complex mental system called the 'self-memory-system', which consists of the interaction between a 'working self' and a 'long-term self'. The working self is task-driven and focused on short-term goals. It exchanges information with an episodic memory system that operates largely out of awareness and produces sensory-perceptual-affective reconstructions of past experience. These transitory memory images are available for use by the working self for possible integration into the long-term self. The latter consists of two components: the 'autobiographical knowledge base' and the 'conceptual self'. The autobiographical knowledge base is a hierarchical structure that stores past experience at increasing levels of abstraction: general events, lifetime periods and the life story schema. Within each of these slots, one can search for and retrieve more summarized or specific autobiographical memories that may be linked to more specific memory images from the episodic memory system.

The other component of the long-term self, the conceptual self, consists of 'socially-constructed schemas and categories that help to define the self, other people, and typical interactions with others and

the surrounding world'—personal scripts, possible selves, self-with-other units, conceptual aspects of internal working models (see Sect. 2.4), relational schemas, self-guides, attitudes, values and beliefs (Conway et al., 2004, p. 500). All this abstracted knowledge about the self is contextualized in terms of a person's life by autobiographical knowledge and ultimately grounded in episodic memories of specific experiences.[21]

The point here is that episodic memories are durably retained only if they have become linked to conceptual autobiographical knowledge; otherwise, they are rapidly forgotten. In the self-memory-system, therefore, the notion of autobiographical memory is no longer defined in terms of episodic memory; it denotes 'a store of information a person possesses about herself, of which episodic memory is only one possible aspect or instance' (Hoerl, 2007, p. 637, n. 4). It is the *conceptual* organization of episodic memories within the self-memory system that transforms them into autobiographical memory and allows them to play a role in constructing and maintaining a coherent, stable mental representation of the self over time.

In this perspective, in contrast to Tulving's assumption that remembering past events serves to establish the sense of continuity of our self over time by virtue of a specific phenomenal quality (i.e., the immediate feeling that 'I' experienced the remembered event), Conway proposes the opposite: it is the conceptual self (the present Me) that selects and also distorts personal memories so as to increase the sense of personal continuity. As a consequence, self-continuity is not 'provided by the identity of the remembering I, but by the perceived similarity of the present and past Me' (Habermas & Köber, 2015a, p. 153).

In conclusion, phenomenological continuity cannot be 'theoretically and empirically separable' from narrative continuity, as Prebble et al. claim (2013, p. 818). Our experience of selfhood across time *is* our feeling of being here as being here in a certain way, through representing to oneself one's own person as a person of a certain type (see Sect. 3.4).

[21] 'Thus, an individual who held a view of himself as 'practical' instead of 'intellectual' might have a lifetime period representation of his time at university as being largely negative. General event and specific episodic memories might be preferentially available to confirm this belief' (Conway et al., 2004, p. 500).

4.3.1 Dissociation of the Jamesian Selves

It is important to notice that the transition from a bodily and social identity to a subjective identity is not an all-or-nothing matter.

To investigate the development of a temporally extended sense of self, Daniel Povinelli and his collaborators (see Povinelli, 2001 for a review) developed a series of experiments based on a variation of the classic mirror task. In the delayed video self-recognition paradigm, the experimenter is filmed surreptitiously placing a large sticker on the participant's head, during a distractor task. The sticker remains on the participant's head for a period of three minutes, after which time the participant views the original video recording of the sticker placement. Reaching up to remove the sticker from one's head after viewing this recording is taken to indicate the possession of a temporally extended self-representation. The logic here is that only if the participant recognizes that the individual in the recording of this earlier event is the same individual watching the recording in the present will they recognize that the sticker is on their head here and now, and hence reach up to remove it. In typical development, this task is passed from around 4 years of age.

Povinelli's (2001) interpretation of these findings is that the concept of a temporally extended self emerges at around 4 years as a function of domain-general changes in the child's representational capacities. Following Perner (1991), Povinelli argues that at about 18–24 months of age infants are able to hold in mind a single representation of an event or object (including one's self) while their perceptual system engages with a primary representation (i.e., current reality). This early system of self-representation underlies the capacity to recognize one's self in the mirror: the infants are able to construct and hold in mind a (secondary) representation of the self while they, at the same time, attend to the image reflected in the mirror (a primary representation of the self) and set up a relation between the two. At about 4 years of age, however, children become able to pass the delayed video self-recognition test because they developed the ability to simultaneously entertain various conflicting representations of the same object or event. This ability enables them to hold in mind, at the same time, various conflicting secondary representations

of the self and to understand the causal connection between past, present and future self-states. The new representational ability, therefore, makes possible the emergence of 'an abstract historical-causal self-concept […] which integrates memories of previously unrelated states of the self into an organized, coherent, and unified autobiographical self-representation' (Fonagy et al., 2002, p. 247).

It can be doubted, however, whether the delayed self-recognition measure is evidence of the emergence of a continuous *psychological* self through time. For if the task is a valid measure of self-awareness as a psychological self, ASD subjects should perform badly on it. But they do not: ASD children can recognize themselves in the delayed image as effectively as do 4–5-year-old typically developing children (see Dunphy-Lelii & Wellman, 2012; Lind, 2010). This suggests that recognizing oneself in the delayed video is really evidence of the capacity to establish causal and temporal relations between past and present states of the self, but that the self in question is the *physical* self and not the psychological one. ASD subjects, then, appear to possess a coherent representation of their own bodies across time; however, being impaired in mindreading abilities, they cannot make the transition from physical to introspective self-description. As Williams puts it,

> …contrary to [the] theory that awareness of the physical self and awareness of the psychological self each depend on the same underlying representational system, these findings suggest that each is underpinned by its own dedicated system, only one of which is impaired in autism. Individuals with autism appear to possess a coherent representation of their own bodies (even across time), despite failing to recognize aspects of their psychological selves. (Williams, 2010, p. 486)

* * *

The hypothesis of a dissociation between bodily and psychological aspects of self-consciousness is congruent with data from cultural psychology and ethnopsychiatry that show that the predominant self-consciousness in adults in preliterate cultures is primarily physical and social rather than psychological. Semi-literate or illiterate adult subjects in preindustrial cultures show an insufficient capacity to represent a virtual

inner space of the mind. In this psychological-cultural condition, dreams (as can be observed in children under 3 years old—see Meyer & Shore, 2001; Piaget, 1929) are not conceptualized as the product of one's own mind, but rather as visions originating from the outside; emotions and passions, being experienced as objective rather than subjective events, are directly ascribed to chance accidents of the body, or are perceived as the effect of 'being possessed' by some force or entity that comes from the outside; thinking is confused with speaking (here 'I think' essentially means 'I say' or 'I tell myself'); furthermore, plans and fantasies are only partially objectified, and hence examined with difficulty. In any case, all these events are always discontinuous, that is, unrelated with each other, insofar as they are not causally integrated within a unitary inner space. The individual feels only partially responsible for them.

Early evidence concerning the difficulties of illiterate subjects in representing an inner experiential space was uncovered by Luria (1976) during two expeditions to Central Asia in the early 1930s. Luria's perspective was that of the historical-cultural school, and in this theoretical framework the construction of introspective self-consciousness requires that our species' neurocognitive mechanisms be accompanied by a collection of conceptual and (indissolubly) lexical tools, of an abstract kind. Where these tools are deficient, as occurs in preindustrial cultures, great difficulty in reflexively and objectively representing a virtual inner world may be observed.[22] One factor responsible for such a deficiency may be the predominance in these social contexts of a 'practical' intelligence rather than an 'analytic' one (Sternberg, 2012).

The exclusive appeal to a practical form of intelligence gives rise to a subjective sphere that fosters somatic-pragmatic rather than psychological conceptions of the individual. Consequently agents conceive of themselves essentially in terms of physical identity, and it is physical identity that forges social identity. The agent then considers herself responsible insofar as she is held *socially* responsible for her actions, whether they be past, present or future. By contrast, the agent is never fully able to responsibly and self-critically appropriate the products of

[22] Luria's findings were later replicated in Western Africa by Michael Cole, Sylvia Scribner and their colleagues. See Cole, Gay, Glick, & Sharp (1971), Scribner & Cole (1981).

her own mind, given her difficulty in constructing an inner experiential space. In William James' idiom, these subjects possess a material self and a social self but lack a spiritual self. All acts (including linguistic ones) are certainly 'produced,' and the agent considers herself as the owner of these acts, as the body clearly identifies their origin and continuity; yet dreams, fantasies, plans, passions, anxieties, frenzies and sorrows, can be identified and conceptualized only with difficulty because their origin and phenomenological place are unclear. Nor, consequently, can we find a full conceptualization of intentionality, as it is expressed not only in emotions but also in fantasies and plans. In such cases, therefore, fantasies and plans are always confused one for the other, since their respective origin can never be traced. All this implies a series of important and serious limitations both in planning future activities and in evaluating past ones.

In these psychological-cultural circumstances, it is quite consequent that the agent evinces a hysterical splitting tendency, which is psychological, but also ethical. The subject 'disclaims' her action (which consists in the body's moving or paralyzing), and the psychic state, being experienced as an objective rather than subjective event (i.e., something that is not produced by the mind but which 'happens'), is ascribed to chance accidents of the body. This inclination to 'acting out' was recorded by Jervis (2011, Chap. 4) in the context of a team study of the ecstatic healing cult of tarantism in the Salentine Peninsula of southern Italy at the end of the 1950s (de Martino, 2005). The psychiatrist noticed that illiterate farmers exhibited a clear tendency to somatize symptoms of anxiety or depression into complex psychoesthesias or hysterical dysfunctions of some part of the body. Owing to their objectivation in some part of the organism, all these symptoms were

> …almost ejected out of the subject's personality and in any case expelled from the sphere of voluntary control; the mental disorder took on a bodily representation, which was often symbolic and not infrequently in contrast with logic; it was experienced in the body, acted outside, and mimed in a showy restlessness or in a helpless motionlessness with no apparent justification. (Jervis, 2011, p. 115)

If objectifying the disorder in the body is not possible, the subject may split from herself a part of the responsibility for her own acts and project such a part onto some force or entity coming from the outside. This is the case with the possession syndrome which is a part of the psychopathology of acute hysterical dissociation (see Jervis, 1969). Possession as hysterical manifestation is a syndrome that borrows specific traditional elements (such as demons, spirits, saints) from the ideological and cultural repertoire of the environment in order to give shape to a number of manifestations which are always characterized by the fact that their causes are objectified externally, thus freeing the subject almost completely from any responsibility for her actions.

These socially constituted patterns of acting out are characterized by Averill (1980) as 'disclaimed action emotions'. A disclaimed action emotion is 'a transitory social role' that is 'interpreted as a passion rather than as an action' (p. 312). These social roles are *transitory* because individuals play them exclusively in short-lived and stressful situations. They allow a behavior that would be unacceptable in other circumstances—that is, in these cases the passive character that is ordinarily ascribed to strong emotions and to sudden passions (love or aggressive) is exploited to avoid responsibility for the action. Moreover, such roles are *covert* in the sense that they take shape only in so far as society does not recognize either their function or the social practices including these roles. Culture-bound syndromes such as running amok or 'wild man' syndrome are cases of disclaimed actions modeled on emotion.

Some similarities can be identified between the typical rationalizations of members of the above mentioned preliterate communities and aspects of the ideology of passions of archaic Greek civilization described in the classic work by Eric Dodds (2004). For Greeks of the archaic period the experience of passions was a mysterious and unsettling event in which the individual felt a force that was inside him—a force which he did not possess, but rather was possessed by. For example, in Homeric poems the experience of divine temptation or infatuation (*ate*) is an outer, objective force that takes possession of the mind, clouds one's consciousness and temporarily makes one mad.[23]

[23] 'Not I', Agamemnon declares, 'not I was the cause of this act, but Zeus and my portion and the Erinys who walks in darkness: they it was who in the assembly put wild *ate* in my understanding, on that day when I arbitrarily took Achilles' prize from him. So what could I do? Deity will always have its way' (*Iliad*, 19, 86ff.; cit. in Dodds, 2004, p. 3).

To explain this tendency to represent mental events as entities originating from the outside, Dodds invokes the concept of 'shame-culture' developed by Margaret Mead and Ruth Benedict. Homeric society rested on the value of *τιμή* (i.e., 'honor', 'public esteem'), and this caused the agent to feel a sense of shame for those aspects of his conduct for which he was blamed by the community. In this sociocultural context, therefore, individuals had a strong need to shift onto outer agents (gods or impersonal entities) the responsibility for behavior that was considered unacceptable. With his seminal studies, therefore, Dodds offered a reconstruction of an intermediate cultural and historical phase, characterized by the objectivation and autonomization of passions from the bodily experience, but still without a full conceptualization of the space of the mind. Here is the example of a significant phase of cultural transition from the primitive difficulty to conceive of the subjective or 'inner space' dimension to the modern conception of consciousness and interiority.

We suppose, then, that the social universe of preliterate people living in preindustrial cultures may foster a prevalently practical intelligence, lacking the necessary resources to make the complete shift from a physical to a psychological form of self-consciousness: the self-consciousness of 'ourselves' as educated members of industrialized societies. But of course this distinction between illiterate and educated people is not clear-cut. First of all, it is easy to note that even our assumption of responsibility for passions or moods that we ourselves produce is often incomplete. For it is not unusual that, in the face of responsibility for committing a serious offence with full lucidity, even educated subjects take refuge in splitting mechanisms. The ordinary verbalization 'I was out of my mind' (as one may say, conforming to a sort of rhetoric, 'I was out of my mind because I was blinded by rage' or 'by passion') easily turns into 'I was not myself', and even into 'Something inside me acted'—from here it is but a short step to hysteria. Moreover, an inclination to consider dreams as a form of access to a world that is not intrapsychic and individual (i.e., produced by the subject), but rather impersonal or transindividual, is also pervasive in our culture. For many people the symbols of dreams are 'already there', deposited in an arcane dimension from which the subject draws during the sleep.

That being said, it is also important to take note of Anthony Giddens' claim that the 'post-traditional' settings emerging from modernity's dynamism have contributed to forge persons marked by *a heightened level of self-reflection*. Giddens argues that late modernity is a 'post-traditional order, in which the question, 'How shall I live?' has to be answered in day-to-day decisions about how to behave, what to wear and what to eat—and many other things—as well as interpreted within the temporal unfolding of self-identity' (Giddens, 1991, p. 14). In other words, the late modern self is 'a reflexive project—a more or less continuous interrogation of past, present and future' (Giddens, 1992, p. 30).

This leads us to the next section, where we will examine 'the self as reflexively understood by the person in terms of her or his biography' (Giddens, 1991, p. 53).

4.3.2 The Thread of Life

The construction of psychological self-consciousness evolves, therefore, within an interplay of mentalization, autobiographical memory and socio-communicative skills modulated by cultural variables. At 3–4 years of age the child turns her mindreading capacities upon herself under the influence of mind-related talk from caregivers; at 5–6 years of age she begins to understand that *psychological* states persist through time and influence current behavior (see Lagattuta, 2014). It is then that the child begins to grasp her subjective identity in terms of autobiography: she begins to integrate memories of previously unrelated psychological states of the self into a coherent causal-temporal organization around a self-concept extended in time. This organized, coherent and unified autobiographical self-representation is *narrative identity*.

Over the last three decades Dan McAdams has developed a theory of narrative identity at the interface of personality psychology, life-span developmental studies and cultural psychology. Building on James' I/Me distinction, Erikson's view of identity and the tradition of the Study of Lives, McAdams (1985) proposed a theory of identity development in which narrative identity is seen as a cognitive structure designed to provide that sense of temporal sameness and continuity that Erikson thought to be a defining feature of identity. Around the same time, Katherine

Nelson (1989) proposed a theory of early narrative development that has since been associated to McAdams' theory (see for references, McLean & Syed, 2015, p. 2).

According to McAdams, narrative identity is a person's life story, the broad narrative of the Me that the I(-ing) composes, edits, and continues to work on. This autobiographical narrative is aimed at providing the jumble of autobiographical memories with 'some semblance of unity, purpose, and meaning' (McAdams & Olson, 2010, p. 527). In other words, people make sense of their own lives by making the Me into an internalized drama, complete with setting, scenes, characters, plots and themes.

The developmental origins of narrative identity lie in the emergence of autobiographical memory in early childhood and the development of *autobiographical reasoning* skills in late childhood through adolescence. Autobiographical reasoning is 'a process of thinking or speaking that links distant elements of one's life to each other and to the self in an attempt to relate the present self to one's personal past and future' (Habermas & Köber, 2015b, p. 3; see also Habermas, 2011). The social-cognitive competencies underlying such a process include the ability to put past events in temporal order (temporal coherence), the ability to account for changes or developments in the self over time (causal-motivational coherence), and the ability to summarize and interpret themes within stories and to apply these to one's own life (thematic coherence).

Habermas and de Silveira's (2008) study showed that a life narrative begins to emerge in middle childhood, but the coherence of this narrative (in all its three dimensions) increases during adolescence. Participants from age 8 through 20 were asked to narrate seven personally significant events and then to place them on a personal timeline. Although the 8-year-olds scored above chance on this task, it was not until age 12 that children began to link single events causally, and causal and biographical reasoning used increased in complexity and coherence across age (see also Reese, Yan, Jack, & Hayne, 2010). Köber, Schmiedek, and Habermas (2015) longitudinally extended this study to explore the development of global coherence in life narratives from childhood to adulthood. It was found that measures of temporal and causal-motivational coherence increase substantially across adolescence up to early adulthood, as does thematic coherence, which continues to develop throughout middle adulthood.

Importantly, Habermas and Köber (2015a, 2015b) argue that autobiographical reasoning is *constitutive* of narrative identity. It embeds personal memories in a culturally, temporally, causally and thematically coherent life story; thus, the life story format establishes and re-establishes the diachronic continuity of the self. More specifically, autobiographical reasoning is a mechanism to compensate for threats of self-discontinuity. In times of relative stability, self-continuity can be established by the mechanism suggested by Conway: remembered self is systematically distorted by automatically assimilating it to the present self-concept, increasing the similarity between the present and remembered reflected self, in order to maintain conceptual self-sameness (see Sect. 4.3). Such a mechanism, however, 'does not provide a mechanism to create self-continuity when *change* is acknowledged' (Habermas & Köber, 2015a, p. 155). In times of biographical change and rupture, self-continuity can be buffered by autobiographical reasoning, that is, the use of arguments that bridge change by embedding it in a larger life story context.

* * *

The claim that full-fledged psychological self-awareness is constituted by means of a life story through which one achieves diachronic unity is an empirical thesis about the development of the self. A *narrative* account of personal identity elaborates this empirical thesis in a claim about *practical* identity (i.e., personal identity considered in its connection to ethical concerns, as in Locke's theory of person). The claim is that we constitute ourselves as *persons* (i.e., as morally responsible agents) by forming and using autobiographical narratives. The unity of a person is thus a particular kind of psychological unity: the unity of an autobiographical narrative (see DeGrazia, 2005; MacIntyre, 1984; Schechtman, 1996; Taylor, 1989).

In some cases, narrative accounts of personal identity are characterized in opposition to what has been, at least until quite recently, the most popular view of personal identity: a significantly amended version of Locke's relational memory criterion (see Shoemaker, 2016). Here the question is one of *reidentification*: what makes a person at time t_2 the same person as a person at time t_1? But when the focus shifts from solely metaphysical puzzles about the persistence of complex objects (such as the ship of Theseus) to the relation between identity and practical and evaluative concerns, the question becomes one of *characterization*:

> …which beliefs, values, desires, and other psychological features make someone the person she is. The reidentification question thus concerns the logical relation of identity, whereas the characterization question concerns identity in the sense of what is generally called, following Erikson, an 'identity crisis'. (Schechtman, 1996, p. 2)

According to the proponents of the narrative view, an answer to the question of characterization requires an acknowledgement that there is more to our personal identity than mere psychological continuity. The mere fact that person *A* at an earlier time t_1 and person *B* at a later time t_2 are psychologically connected does not entail the presence of the activities of *self-interpretation* and *self-creation* that are central to our experience of being persons. And what enables persons to be actively self-interpreting and self-creating agents, is the construction of self-narratives (see Mackenzie, 2008; Schroer & Schroer, 2014).

Now, we agree that typical accounts of personal identity which emphasize psychological continuity fail to capture the activities of self-interpretation and self-creation central to the experience of being a person and to personhood more generally. Yet what, in this context, is 'self-interpretation'? Authors such as Alisdair Macintyre and Charles Taylor view the self as a self-interpreting being in a sense inspired by the hermeneutical tradition. In this tradition, the self 'is not a thing; it is not something fixed and unchangeable, but rather something evolving, something that is realized through one's projects, and therefore something which cannot be understood independently of one's own self-interpretation' (Zahavi, 2003, p. 59). There is much here that is worth developing. However, we should be cautious about ideas from a philosophical tradition that is foreign to our naturalistic commitments. Building on recent theoretical systematizations in personality psychology, we have defined narrative identity as the ability to construct an internalized and evolving story of the self that can provide a life with some degree of meaning, unity, and purpose. And in our view, such an active process of self-interpretation is a theory-driven narrative re-appropriation of the products of the neurocognitive unconscious. The self is thus a self-interpreting being in a naturalistic sense that is fundamentally foreign to the hermeneutical tradition. A hermeneutical notion of self-interpretation, emphasizing meaning at the expense of the

psychobiological theme of the unconscious, risks, as already argued in Sect. 2.2, in surreptitiously reintroducing the idealistic conception of the conscious subject as primary subject.

Things are more complicated in the case of Paul Ricoeur, another standard reference in the literature on the hermeneutical view of narrative self. In *De l'interprétation*: *essai sur Freud* Ricoeur made an important attempt to overcome the contrast between 'energetics' (i.e., Freud's drive theory) and hermeneutics. He investigates how psychoanalysis allows for both the hermeneutical theme of meaning and intentionality and the objective and biological theme of drive causality. Within this framework, Ricoeur debunks the versions of psychoanalytic interpretation which are unilaterally aimed at the subjective or intersubjective 'reconstruction of meaning', in keeping with the standards of interpretive conventionalism. This attempt of synthesis, however, remains within a conception of the unconscious that we have rejected. Ricoeur coins the term 'anti-phenomenology' to define Freud's inquiry into the unconscious. This inquiry is characterized as 'an epoché in reverse' because 'what is initially best known, the conscious, is suspended and becomes the least known' (Ricoeur, 1970, p. 118). Consequently, whereas the phenomenological tradition pursues a reduction of phenomena to consciousness, Freud's methodological approach aims at a reduction of consciousness: the latter loses the Cartesian character of first and last certainty, which stops the chain of methodical doubts on the real, and becomes itself an object of doubt. Thus Freud's psychoanalysis becomes a 'demystifying hermeneutics' (see Sect. 2.2, note 17). However, as we have seen above, in reality Freud's inquiry into the unconscious really starts from consciousness taken as *given*; and this makes psychoanalysis a dialectical variant of phenomenology. In contrast, a dynamic psychology informed by cognitive sciences is not vulnerable to this objection: for it aims to pick up the critical content of Freud's psychoanalysis, its being a demystifying project, but within a framework where the unconscious is understood in that subpersonal/personal dialectic which we championed in Sect. 2.3.2.

Certainly, even if we describe, naturalistically, self-interpretation as a re-appropriation of the products of the neurocognitive unconscious, self-narratives are *not* merely the result of the development of a psychobiological system. In the previous section we saw how sociocultural

variables may significantly modulate the construction of psychological self-consciousness. A socially and historically situated narrative unity is needed to fashion that form of inner life characterized by the heightened level of self-reflection to which Giddens refers.

Yet, whereas the narrative view of personal identity makes the socially and historically situated narrative self *the* foundational aspect of human selfhood, we have argued that the narrative self is certainly social and historical but also, and perhaps still more, biological and psychological, since the contribution of society and history has very clear limits, fixed by the structure of the psychobiological system. From this point of view, there is some similarity between our account and Dennett's naturalistic narrativism (see Sect. 3.1.2). However, as we will see in the next chapter (see Sect. 5.4), there is also a crucial difference concerning the 'degree of reality' of narrative identity, and its causal role.

5

The Self as a Causal Center of Gravity

In the previous chapter, we drew on developmental, dynamic, social and personality psychology to put forward a view of the onset of self-consciousness as the establishment of a process of self-description that is a unifying, integrative, synthesizing selfing process. We concluded by distancing ourselves from the non-naturalistic strands in the hermeneutical conception of narrative identity. Now we aim to show how our approach to the self enables us also to reject the antirealist argument that infers, from the non-primary, derivative nature of the self, a view of it as an epiphenomenal by-product of neurobiological events or, alternatively, of social (or sociolinguistic) practices. The antirealists—we will argue—disregard the inherently *defensive* nature of identity self-construction. This psychodynamic component has hardly been noted by the philosophers who have made use of cognitive sciences to put forward a theory of the self. Yet we argue that it is precisely such a component that is the keystone of a philosophical anthropology congruent with the findings of the psychological and brain sciences.

Defenses can be explained only by placing them within a more general theme: that of the *fragility of the subject*. This theme will be articulated by

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_5

reference to three concepts: Freud's 'fragility of the ego', Ronald Laing's 'ontological insecurity' and Ernesto de Martino's 'presence'. The picture that emerges is the following.

Self-consciousness as subjective identity means finding oneself at the center of one's own orderly and meaningful subjective world, hence at the center of a historical and cultural environment to which one feels one belongs. But this full-fledged psychological self-awareness is a *precarious* acquisition, continuously constructed by the human subject and constantly exposed to the risk of falling apart. This precariousness allows us to grasp the intrinsically defensive nature of subjective identity. The need to construct and protect the most valid identity possible is rooted in the subject's primary need to subsist subjectively, and thus to exist solidly as a describable ego, as a unitary subject. The incessant construction and reconstruction of an acceptable and adaptively functioning identity is the process that produces our intra- and interpersonal balances, and thus must be regarded as the foundation of psychological well-being and mental health. The selfing process, therefore, imposes a teleology of self-defense on the human psychobiological system. Thus subjective identity (the self that selfing creates) is not an epiphenomenon but a layer of personality that serves as a causal center of gravity in the history of the system.

5.1 A Baconian Approach to Defense Mechanisms

First, we need to understand the notion of defense mechanism.

Freud developed the theory of the 'defense mechanisms' mainly in the 1926 essay *Inhibitions, Symptoms, and Anxiety*, and it would later be fully worked out in the works of Anna Freud and Melanie Klein. However, for those who, like us, aim to include dynamic psychology within the theoretical and methodological framework of cognitive sciences, the way in which Freud and many of his successors in the psychodynamic tradition have dealt with the study of psychological defenses must undergo a radical revision.

In Sect. 2.2, we said that the Freudian conception of the unconscious is limited by an insufficient emancipation from Descartes' model of the *passions de l'âme*. A general epistemological implication arises from such a model: rational consciousness cannot *per se* be mistaken, and thus error does not arise within such consciousness but is due to the influence on the mind of the passions, which are the emotional, visceral, impulsive-instinctual, 'animal' motions originating in the opacity of the bodily machinery. Here is the thing: Descartes' faith in reason as a producer of truth, the idea that what is clear and distinct cannot be false, and that errors are essentially a sort of derailment due to drive-visceral interferences, is implicit in Freud's idea of a fundamental conflict between visceral and rational influences, which are at the root of the psychic drama and should result in a victory of the latter over the former.

It is to be noted, however, that the Cartesian conception of error had already found an implicit refutation in the approach that Francis Bacon expressed in the *Novum Organum*. Bacon, unlike Descartes, does not think that the errors of judgment and conduct can be traced back to affective influences on rational consciousness. Rather, he thinks that the conscious and rational mind is a 'crooked mirror' that naturally produces errors.[1] The famous 'idols', constant factors of deception, are, according to Bacon, human knowledge's habitual way of operating. In current terms, this philosopher sees the mind's errors, illusions, and self-deceptions as intrinsic to the ordinary cognitive processes. It is on these grounds that Bacon claims the necessity of a system of tests through which our spontaneous tendency to make errors is rooted out and rectified by the method of research, on the basis of a rigorously empiristic methodological principle (see Jervis, 1984, Chap. 3; Rossi, 1968).

Today the cognitive science research work on emotion and thought is largely Baconian, providing us with the tools to deconstruct the ideology of the conflict between reason and the passions, from which Freud failed to extricate himself. On the one hand, the phenomena that folk psychology labels as 'emotional' can no longer be relegated, as the ideology of the passions suggested, to a 'low' and 'primitive' psychic sphere which threatens the

[1] 'The human understanding is like an uneven mirror receiving rays from things and merging its own nature with the nature of things, which thus distorts and corrupts it' (Bacon, 2000, p. 41).

nobility of 'the thinking thing'. All such phenomena belong to the wider universe of *all* mental events (see Griffiths, 1997, 2004b). On the other hand, the experimental investigation of rationality and reasoning shows that in this case, as in the case of emotions, there is no unitary cognitive sphere; the factors of error are seen as inherent in rationality, or rather immanent in that hodgepodge of procedures and abilities into which our bounded rationality can be decomposed (see Carruthers, 2014; Stich, 1990).

This leads us to a radically new interpretation of the Freudian idea that self-consciousness is a construction packed with self-deceptions and bad faith. For, in the Baconian perspective, the aspects of ambiguity, self-deception and suffering of human life can no longer be conceived as much of the philosophical tradition has viewed them, namely, as the crisis of a fundamentally rational agent, temporarily overwhelmed by the perturbing influence of affects and sentiments. These aspects can now be conceived as *globally constitutive* dimensions of the mind and conduct.[2] This gives place to a reinforcing overturning of the psychoanalytic questioning about defenses: what we now have to ask ourselves is not how and why some defensive mechanisms exist, but rather if it is not the case that *the structures of knowledge and action around which everyday life is built serve defensive functions* (see Jervis, 1993, pp. 301–302).

To put it another way, defense mechanisms are mechanisms that permit us to think and act. Although their most manifest function is that of protecting us from anxiety, defense mechanisms are the primary instruments for creating order in the mind. Consequently, we grasp something that is already present in Freud but which the Cartesian framework prevented him from articulating fully: the defensive processes are much more than bulwarks against anxieties and insecurities that perturb the order of our inner life; they are the primary instruments for establishing order in the mind; they are the very structure of the mind—the Freudian ego itself is the defense system.

* * *

[2] Thus, for example, self-deception can no longer be conceived as a pathology of belief-formation, the temporary crisis of a fundamentally rational agent, which can be explained only in terms of a non-rational psychological sphere that can be clearly demarcated from the workings of our self-conscious rationality. Now self-deception is a natural inclination of the human mind, a property inherent to belief-formation mechanisms.

Within this theoretical framework, dynamic psychology joins forces with interpersonal and social psychology. The defense of self-image, the social attitudes in general and the stereotypes and prejudices in particular—along with the rationalizing management of cognitive dissonance—are the building blocks of an interpersonal and social reality which are packed with systematic errors, or, as Freud would have put it, interested self-deceptions. And all these structures of self-deception are defensive constructions that spring from mental operations in which the cognitive aspect cannot be separated from the affective. To illustrate, we briefly focus on the social psychology of prejudice.

Within the framework of the systemic naturalism outlined in Sect. 2.4, 'knowing', as well as 'making sense', is primarily a pragmatic matter, a 'knowing how to do things'. In the context of everyday life, an object makes sense for me and is known by me, because I place it in a pragmatic context, insofar as I consider it within a repertoire of competences: I have done something with this object in the past and I can do something with it in the future. But, inherent in the very idea of 'knowing how to do', there is an organization of the world according to differentiations and hierarchies. All of us, in forming more or less complex behavioral patterns, act according to gradients of involvement and interest. Basically, we assign different values to single objects and to different aspects of our behavior itself.

'Values' are to be understood here as simple differences of importance, that is, of priority, in the context of the general theme of adaptation. There is an objectivity in the gradients of value in specific contexts. In the cycle of everyday activities, animals organize their behavior as a function of a limited series of general interests (i.e., evolutionary values) such as avoiding predators, foraging, selecting fertile mates, deterring sexual rivals, negotiating dominant hierarchies, managing aggression, detecting alliances and so forth. Each of these general needs dominates over specific behavioral patterns which, from time to time, represent a higher priority than others, that is, they literally 'come before' insofar as they 'have more value', alternating with each other at the top of the agenda of 'things to do'.[3]

[3] Such behavioral priorities can be quantified by means of game theory. See Maynard-Smith (1982).

In our species, too, no object can be perceived or used outside of such appreciations. The panorama of reality, thus, takes shape in accordance with our interests in objects, namely, according to the value that we assign to our surroundings:

> Clusters, hierarchies of values arise; the various areas of reality are assigned different grades of importance. The 'nearer' scenarios are those that we are more interested in, and are more easily the object of our 'positive' planning; the more 'distant' scenarios are those we are less interested in; they are less differentiated in their internal details, and can more easily appear to be extraneous or even hostile. (Jervis, 1993, p. 331)

These variables come to be organized in the first place, according to the phenomenological category of *domesticity* (or *familiarity*):

> All of us tend to make a spontaneous separation between, on the one hand, what is 'internal' to a limited, 'domestic' social world, and hence 'good' and 'reassuring', where we find, as it were, a proximal panorama of guaranteed values; and, on the other, what is 'external', 'alien', which we are less interested in, whose guaranteed value is lower, and where objects and events can take on negative tones. (Ibid.)

This way of organizing reality, and of situating ourselves at its center, is constrained by invariant psychological structures. It is a primary way of establishing order which has clear affinities with some basic structuring categories such as 'before-after', 'high-low' and, above all in our case, 'inner-outer' and 'near-distant'. The phenomenological category of domesticity refers to the experience of the world environment as structured according to criteria of distance and controllability. This is a primarily cognitive operation, but one which is nevertheless linked to the attribution of emotional-evaluative connotations in conformity with the above-discussed core affects. As mentioned in Sect. 4.2.2, valence is an elemental, binary, antinomic dimension of the agent's dispositional orientation toward reality which sharply distinguishes between good and bad, friend and foe, and thereby 'approaching/withdrawing', 'accepting/rejecting', 'incorporating/expelling'.

In animals, the world tends to be organized in accordance with the category of territoriality; in ways that differ across species, we find the den to be the most protected shelter, followed, moving outwardly, by a 'possession zone', then an 'exploratory zone', and so on. In infants, the 'domestic space' is linked to the presence of the primary attachment figure; as described in Sect. 2.4, the possibility of exploring, leaving the 'protection zone', is proportional to the level of reassurance provided by the caregiver. In adults the difficulty of leaving the 'domestic zone' has been called 'territorial anguish' by de Martino (1951–1952), and viewed by the philosopher-ethnologist as one of the two main parameters of the feeling of being in crisis: the spatial or geographic parameter as opposed to the temporal (see Sect. 5.3).

This brings us to prejudice, as its psychological dynamic belongs precisely to the way of organizing reality and placing ourselves at its center that we have just sketched (see Jervis, 1996, p. 774). That is, the dynamics of prejudice are part and parcel of the ways in which we spontaneously systematize material or social reality according to categories of relevance and gradients of approval and disapproval. The peculiarity of prejudice consists in the fact that, whereas in most of our basic attitudes (of liking, curiosity, identification, wishing, disposition to the affective bond, etc.) there is a ('positive') tendency to approach the object, in prejudice we find the opposite tendency to reject the object, resulting in a refusal to know it.

Evolutionary, developmental and social psychology suggest that people are predisposed to categorize the self and others into groups. Any minimal grouping, based on race, ethnicity, nationality, religion or arbitrary assignment, tends to produce a preference for the in-group (us) over the out-group (them) (see Kinzler & Spelke, 2007). Now, according to the social identity theory paradigm (see Tajfel & Turner, 1986), the dynamics of feeling a member of the ingroup is closely linked to stigmatizing the outgroup members as treacherous and different. As a result, the sentence expressing the prejudice (i.e., the stereotype) at the moment at which it brings discredit on 'the others', accomplishes the defensive (self-apologetic) function of enhancing our self-image, providing us with a collective identity (a sense of community), which is also a certificate of nobility that 'the others' do not possess. Feeling comfortably part of a 'valid' community causes us to believe in our inner validity.

Thus the biasing aspect of prejudice can be ascribed to the very ways in which our knowledge of social world constitutes itself.

* * *

To recapitulate, we have argued that the Freudian view of defense mechanisms must today be subjected to a radical revision, as a result of a paradigm shift. Freud's investigation was carried out wholly within a Cartesian logic, where rational consciousness fails only because of the influence of emotional and affective motions originating from the bodily machine. By contrast, some research areas in the psychological sciences adopt a Baconian logic in which error is seen as inherent in the very mechanisms of 'high' cognition. Therefore, whereas in Freud the subject normally deceives herself because she is unable to accept the presence, deep inside herself, of 'inadmissible' sexual and aggressive drives, in a dynamic psychology informed by the renewal of the traditional psychological categories outlined above, the mechanism of self-deception becomes pervasive. This Baconian dynamic psychology, then, takes the form of a systematic study of the mechanisms of self-deception of self-conscious subjectivity; as we have pointed out, these are intrinsically defensive cognitive-affective mechanisms.

5.2 Construction and Defense of Subjective Identity

We are now in a position to identify the ultimate root of the subject's primary defensiveness; it lies in its fundamental *fragility*.

After undergoing a reinforcing reversal, the construct of the defense mechanism (like that of the unconscious) no longer plays the role that Freud assigned to it, namely, that of debunking an idealistic vision of the subject as an entity with a primary identity and force. Much more radically, it certifies the nonexistence of such an entity. What, more than anything else, defines the real human subject is its intrinsic fragility; consequently, what we must try to understand is how, notwithstanding the lack of an identity or a force that guarantees it, the subject is able to construct itself. In other words, the problem is no longer to know how

the subject can 'come down' from the level of nobility at which it was placed, but, on the contrary, how it can 'rise' up to self-consciousness and culture in spite of its 'ontological insubstantiality', and indeed, still more radically, a sort of original 'non-being' (Jervis, 1993, p. 301).

Within this framework, defensiveness is immanent in the selfing process insofar as it consists precisely in dialectically denying the subject's ontological insubstantiality, in mobilizing self-protective measures against the threat of not being there. Or, equivalently, selfing imposes a teleology that is focused on self-defense upon the human psychobiological system.

Such a view will be now argued by discussing developmental, clinical and psychosocial evidence from a psychodynamic standpoint.

* * *

Let us return to McAdams' view of narrative identity as an internalized and evolving story of the self, directed at providing life with some measure of temporal unity and purpose (see Sect. 4.3.2). Within his three-tiered conceptual framework for understanding *personality* (see McAdams, 2013, 2015; McAdams & Olson, 2010; McAdams & Pals, 2006), narrative identity hinges on two other cognitive layers. The first consists of a small set of broad *dispositional traits* implicated in social life (including the so-called 'Big Five': extraversion, neuroticism, agreeableness, conscientiousness and openness to experience) which account for consistencies in behavioral style from one situation to the next and over time. The second layer consists of a wide range of *characteristic adaptations* (including goals, strivings, personal projects, values, interests, defense mechanisms, coping strategies, relational schemata) which capture more socially contextualized and motivational aspects of psychological individuality. Thus, unlike the continuously self-rewriting autobiographies of the Joycean machine, identity as a story of the self is by no means empty chatter; it is the third layer of personality built upon the dispositional base and characteristic adaptations. The process of selfing aims to bring traits, skills, goals, values and experiences into a meaningful life story.

Personality, however, is not only the framework within which self-identity can develop, setting limits to that development and endowing it with its individuality; it is also the premise for the study of the pathologies of identity. The dynamics of the mind, of its anxieties and of its

neurotic discomforts can hardly be ascribed to an accidental disorder; it entails, on the contrary, the study of a personality, with regard not only to its basic features (the first layer) but also to its *balances* and its *defense styles*, which develop and stabilize over the course of thousands of episodes and of the complex affairs of infancy and adolescence.

This point becomes clear when considering dynamic psychology as a research tradition which, breaking with a longstanding philosophical vision that has taken self-consciousness to be a primarily, if not purely, cognitive phenomenon,[4] holds that the construction of affectional bonds and the construction of identity cannot be separated.

In the previous chapter, we saw that during very early childhood, and especially from the third year of life, self-consciousness goes beyond the simple recognition of one's own body to become mentalistic self-description, and later, narrative self-description. This description of the self that the young child feverishly pursues is an 'accepting description', that is, a description that is indissolubly cognitive (as a *definition* of self) and emotional-affectional (as an *acceptance* of self). In practice, therefore, the affective growth and the construction of identity cannot be separated. Children need a clear and consistent capacity to describe themselves in a manner that is fully legitimized by caregivers, socially valid, capable of attracting attention and serving as a base for ceaselessly renewed affectional transactions.

In the course of very early childhood, the development of subjective identity exhibits a paradoxical feature. While each of us, with increasing clarity, constructs and recognizes the singularity of her being herself—a singularity that cannot be confused with others—at the same time and in a contradictory way, everyone 'plays' with identifications, introjections, and projections, intermingling her own personality characteristics, more or less temporarily, with those of others. In the first place, the child's construction of her own identity occurs by means of an introjective appropriation of parts of the identity of others—first and foremost, the idealized characteristics of the parent of the same sex. There is more to it than that, however, for the 'pretend play' so evident from the third year of life, makes explicit the child's propensity to temporarily feel herself

[4] Paradigmatically: 'Self-consciousness is primarily a cognitive, rather than an affective state' (Bermúdez, 2007, p. 456).

different from what she is, to go through fictitious identities, to enhance herself or explore her being and borders, mingling herself with identities that are not hers.

Even adolescent crisis, and together with it the process of social autonomization in post-adolescence, are largely a problem of identity. The most widely referenced model of identity development is still Erikson's psychosocial theory, where identity formation represents the main task of the developmental stage of adolescence.[5] More precisely, the fundamental problem of adolescence lies in discovering how to move beyond the *heteronomy* of identity, through which self-definition has theretofore been linked to the relationship with one's parents. For the adolescent, the problem is now how to perform the risky leap to an *autonomous* self-definition—an identity freed from any protective recognition, mediated by identifications with transitional figures and hinged on non-familial life. In Jamesian terms, the various parts of the material, social and spiritual selves must be organized into 'a new pattern that confers upon the Me a unifying and purposeful sense of identity' (McAdams & Cox, 2010, p. 164). The optimal outcome of this process is a kind of dialectic balance in which the so-called 'syntonic' pole of identity integration is predominant over the 'dystonic' pole of identity diffusion.

It is to be noted that Erikson views identity diffusion as consisting in an insufficient integration of self-images originating from a 'weakness of the ego'.[6] Moreover, in a revision of Erikson's developmental theory, Crawford, Cohen, Johnson, Sneed, & Brook (2004) suggest that young individuals who experience identity diffusion may use cluster B symptoms (borderline, histrionic and narcissistic symptoms) as a form of maladaptive defense against distress, which typically arises from a poorly consolidated identity. Intimacy and engagement imply a constant threat of fusion and consequent loss of a fragile identity, both of which can be defended against by the symptoms and disturbed behavior seen in borderline patients. Finally, psychotic crisis or decompensation, a dramatic and common risk

[5] Since James Marcia's (1966) elaboration of the identity statuses, Erikson (1968) has represented, with some important variants, the dominant approach to the study of identity development (see McLean & Syed, 2015, p. 2).

[6] It is to be noticed that in this context Freud's *das Ich* is taken as a synthetic function, a synthesizing process, and thus coinciding with selfing. See McAdams (1997, p. 57).

between 16 and 18 years of age, can be interpreted to a significant extent as a failure in achieving the autonomy of identity.

In the transition from emerging adulthood to young adulthood, problems with identity have an almost equally prominent role. Individuals who approach their late twenties carrying the burden of mental disorders, disturbed behavior and unresolved social drifts, often begin to suffer acutely from having failed to build an identity that is adult, self-determined, socially recognizable and acceptable; in such an extremely painful crisis, the pre-existing psychological problems can easily worsen. A typical task of psychotherapeutic and psychoanalytic work is the maturational clearing up of infantile remains within the personality; that is, remnants of a lack of psychological autonomy in subjects in their late 20s and, often, early 30s (see Jervis, 1997, p. 74).

Thus, the quest for an identity that is adult, self-determined, socially recognizable and acceptable emerges as a 'center of gravity' of the entire developmental lifespan. A quest that unfolds through the interplay between public sociality and private affectional bonds. Each of us devotes not a small part of our resources to creating situations that guarantee not only material protection but also a positive self-image, and together with it appropriate supplies of self-esteem. In so doing, we seek a confirmation of the solidity of our self-image. In this context, the competition for a social status aimed at providing a suitable 'public' self-image which can guarantee characteristics of an objective 'dignity' for one's own image is connected with the search for more strictly affectional reassurances, namely, with the negotiation of forms of unconditional acceptance from a small number of individuals belonging to the intimate sphere.

Here, though, some possibilities of conflict arise. For example, the conflict between a competitive attitude, aimed at securing a high social status, and a cooperative attitude, with the purpose of securing acceptance and affectional protection. Or, more generally, the contrast between the need for a realistic perception of one's identity and the necessity of pre-

serving self-esteem through a 'high'—and possibly unrealistic—model of ideal identity.[7]

Other more contingent or more strictly individual mechanisms suitable for producing self-esteem and security originate from the ways of self-presentation and the techniques of ongoing management of one's identity in the activities and banal conversations of everyday life. Here Goffman's work (e.g., 1959, 1963) constitutes the essential reference point. Within the context of the conversation, the subject's true purpose in many explanatory or persuasive discourses, rather than explaining facts or convincing an audience of the soundness of a specific practical solution, is actually that of favorably presenting herself (see Antaki, 1981, 1985). Similarly, in the context of self-presentation, the analysis of the ordinary, folk-psychological descriptions and explanations of our own and other people's behavior shows how the defense of the self-image is closely linked to the *self-defensive* use of causal attributions (see Weary & Arkin, 1981a, 1981b).

In the next section, we will see how narcissistic defenses, too, are ordinary strategies aimed at preserving a positive self-image.

In sum, the development of a subjective identity as outlined here lays the ground for a hypothesis about human nature. Human life does not respond only to elementary biological needs such as surviving and reproducing; nor can our motivations be traced back only to universal forms of social competition which can be observed in the rivalry among animals. Rather, our everyday life takes shape in accordance with 'a specifically human necessity', namely, the 'maintenance of identity' (Lichtenstein, 1977, p. 77), an identity that must fulfill a fundamental requirement, that is, it must be 'a self-image endowed with at least a minimal solidity, and that is, solid enough to confirm to ourselves that *we exist without dissolving ourselves*' (Jervis, 1997, p. 33).

* * *

The reference made to Erikson's concept of identity diffusion leads us into the clinical dimension of the inextricable link between identity self-description and self-consciousness. One cannot ascribe concreteness and solidity to one's own self-consciousness if it does not possess

[7] The defensive maneuvers to manage real- and ideal-self discrepancies were first investigated by Duval and Wicklund (1972), on which see Carver (2012).

at its center, and as essence, a description of identity that must be clear and, inextricably, 'good', in the sense of being worthy of love (see Balint, 1965). The incessant construction and reconstruction of an acceptable and adaptively functioning identity is therefore the process through which our intra- and interpersonal balances are produced, and hence, the foundation of psychological well-being and mental health. This finds illustration in the developmental psychopathology of attachment, in which the abusive or seriously neglective behaviors of attachment figures are seen as conducive to disturbances in identity.[8]

Let us return to a point made in Sect. 2.2: the later Freud came to think that that the id reigns over our entire mental life; as a consequence, consciousness lost its importance, and together with it the ego, so much so that he held that the id exploits the ego as a façade. Insofar as the later Freud claimed that the ego is only the façade of the id, his thought contains a theory of the *precariousness* of the ego. But it is important to make it clear that the fragility of the subject cannot really be a central theme in the Freudian theoretical framework. On his view, indeed, the disunity of the mind can be explained not so much in terms of fragility or insufficiency but rather in light of the pervasive nature of unconscious conflicts. It is intrapsychic conflict that is the keystone of Freud's theory of mental disorders.

A new landscape emerges with the psychodynamics of object relations. As early as the 1930s psychoanalysis began to shift its theoretical focus from the affective problems typical of 3–6-year-old children to those of the first year of life, and hence from problems concerning the conflicts stemming from the triangle of Oedipal rivalries to earlier problems arising from a weakness, or fragility, or scarce cohesion, or insufficient integration of those structures of the mind that Freud calls '*das Ich*'. This structural condition of fragility is experienced by the subject as a chronic feeling of insecurity, or lack of self-esteem, lack of confidence in oneself, lack of cohesion of the self (expressions that we take to be essentially synonymous), which cannot be traced back to conflictual neurotic themes.[9]

[8] As Gergely puts it, '…in developmental psychopathology *unrealistically negative dysfunctional self-attributions* are seen to arise from attempts to rationalize the abusive or seriously neglective child-directed behaviors of attachment figures' (2002, p. 42; emphasis added).

[9] It is important to emphasize the distinction between experience and structure: a solid feeling of itself is one of the subjective effects of a solid, cohesive and well-integrated ego, whereas an experi-

Drawing on Ronald Laing's *The Divided Self*, a classic of psychiatric literature, we can describe the experiences originating from a weakness of the ego as symptoms of 'ontological insecurity'. The individual with a firm core of ontological security, Laing says, is one who owns a sense of the self as a cohesive and well-demarcated entity, as well as a consistent feeling of biographical continuity:

> The individual […] may experience his own being as real, alive, whole; as differentiated from the rest of the world in ordinary circumstances so clearly that his identity and autonomy are never in question; as a continuum in time; as having an inner consistency, substantiality, genuineness, and worth; as spatially coextensive with the body; and, usually, as having begun in or around birth and liable to extinction with death. (Laing, 1960, p. 39)

By contrast, the ontologically insecure individual is one who is liable to the collapse of subjectivity described as an experience of disintegration, psychic deadness or numbness, and a sense of moral emptiness: 'He may not possess an over-riding sense of personal consistency or cohesiveness. He may feel more insubstantial than substantial, and unable to assume that the stuff he is made of is genuine, good, valuable' (ibid., p. 42). In the ordinary circumstances of living, such an individual is plagued by a feeling that the living spontaneity of the self has become something dead and lifeless: '[he] may feel more unreal than real; in a literal sense, more dead than alive; precariously differentiated from the rest of the world, so that his identity and autonomy are always in question' (ibid.). Discontinuity in temporal experience is a basic feature of such a condition: 'He may lack the experience of his own temporal continuity' (ibid.). As Giddens notes, time may be understood here as 'a series of discrete moments, each of which severs prior experiences from subsequent ones in such a way that no continuous narrative can be sustained' (1991, p. 53).

Everyday defensive mechanisms are perceived by the ontologically insecure individual as an indispensable bulwark against an outer world and an inner world experienced as threatening. The subject will try to curb this kind of experience by coming up with 'ways of trying to be real,

---

ence of a divided self is one of the experiences deriving from an ego that is fragile and not well integrated.

of keeping himself or others alive, of preserving his identity, in efforts, as he will often put it, to prevent himself losing his self' (Laing, 1960, p. 42). What are to most people everyday happenings 'may become deeply significant in so far as they either contribute to the sustenance of the individual's being or threaten him with non-being' (ibid., p. 43).

* * *

In the context of attachment theory Laing's symptoms of ontological insecurity are seen as the last traces of a remote 'basic fault' (Balint, 1992), which is to be traced back mainly to early deficiencies in the infant-caregiver relationship. Such deficiencies may be the outcome of adverse experiences during the first years of life, such as an extended experience of abandonment (separation or loss), a familial climate of violence, or chronic scarcity of suitable affective attentions by the parents (see, e.g., Kobak, Zajac, & Madsen, 2016).

In this framework, the Freudian metaphor of a fragile and insufficiently integrated ego identifies a condition that predisposes individuals to a broad and varied pathology including, along with psychoses and personality disorders, vaguer conditions characterized by fragility in managing reality and affectional bonds, uncertainty in the use of oneself and feelings of a 'divided self', possibly made more difficult and hidden by conflictual neurotic situations.

At the more serious end of the spectrum thus lies the patient at risk of psychotic disgregation. In patients with schizophrenia, the process of disgregation of the ego functions is in progress, with a consequent loss of the capacity to suitably process information from reality. This results in the destructuration of the experience of external reality and the breaking down of a coherent sense of identity.[10] This twofold destructuration makes the patient unable to clearly discriminate the bounds between the inner space of the mind and the corporeal and extra-corporeal experiential spaces. The subject then develops abnormal defensive measures, aimed at heading off the experiential chaos arising from the process of disgregation.

[10] Thus, schizophrenia is not only associated with an impaired sense of self-agency (see Sect. 4.1.3) but also with disturbance of the temporal dimension of self that assures a stable and coherent sense of identity. See Raffard et al. (2010).

An insufficient solidity of the ego, and the resulting chronic feelings of ontological insecurity plague patients with personality disorders. Let us consider, in the first place, narcissistic personality disorder.

In contemporary psychoanalysis, narcissism is regarded as one of the tenets of the dynamic structure of personality. All the premises for the modern development of this topic are already found in Freud (1914). First, this work puts forward the idea that narcissism is an absolutely necessary phenomenon for establishing a healthy mental life. It is self-love (or self-cathexis), and as such, governs the constitution of the ego; it is the starting point for the subsequent construction of one's identity; it is what prevents the ego from breaking apart. Not only in very early infancy but all through life, self-love constantly restores self-esteem and grounds the capacity to love. Thus, only someone who has a good relationship with herself, accepts herself, loves herself and takes care of herself with affection, possesses the security and wealth of affectional bonds that will spontaneously pour out in work and enthusiasm, in acceptance and willingness.

The most important revision of Freud's theory of narcissism is that of Federn (1952), who argues that schizophrenia is not due to a withdrawal of libido into the ego, as Freud maintained in his theory of 'narcissistic neurosis', but, on the contrary, to a dissolution of ego boundaries because of a *deficiency* of ego cathexis. The more endowed the ego is with narcissistic libido, the more capable it is of holding out against the already mentioned process of psychotic disgregation.

After Freud, Federn makes the most important contribution to the reversal of the traditional view of narcissism: that type of personality that (also without being affected by schizophrenia) appears to us as narcissistic, that is, retreated into its own world, mirroring itself in a grandiose image of its self-sufficiency, is, in reality, *deficient* in narcissism. The narcissist, contrary to appearances, suffers from a chronic deficit of self-love. And it is precisely because he is insecure, does not love himself and believes that he is not worthy of love, that he anxiously watches himself and constantly strives to reassure himself and strengthen his own image—an image that is perceived as fragile, poor and perpetually deficient—with ornaments and illusions. Briefly, narcissism is a normal phenomenon, but becomes pathological when it is exploited to compensate for a condition of insecurity and insufficient self-esteem. Here the fundamental anxiety is *the dread of not existing*.

In the very early infancy there may have been an insufficiency of that narcissistic investment that is normally mediated by a good relationship with the caregiver and in subsequent years no remediation of this insufficiency. The subject is left with what may be called 'a narcissistic credit': the consequent fragility of the ego manifests itself at the simplest level in the development of a personality characterized by insecurity, a deficit of primary self-esteem, low tolerance for frustrations and a continuous need to supply himself (albeit never sufficiently) with self-appreciation and confirmation. The original deficit of self-love and self-esteem, and hence the inner affectional misery, give rise to a difficulty in loving and being loved, come to be connected to poorly managed aggression, and produce a deep self-destructive unhappiness that conceals itself, typically, through complacent, evanescent and grandiose self-illusions, identifications with persons, things or movements provided with the chrism of power, and interpersonal relationships which may at times be handled skillfully but are nevertheless characterized by egocentricity, avidity and instrumentality.

In this context, narcissistic defenses are the ways in which the patient's dramatic struggle to keep herself alive—brought on by ontological insecurity—seeks containment. In other terms, they are the attempts, often sorrowful and at times desperate, to care for and defend one's image as protection for an identity felt as excessively fragile. This theme was explored in depth by Heinz Kohut, who presented afresh Freud's metaphor of the solidity of the ego in terms of the cohesion and self-legitimation of identity.

Note that a narcissistic defense consists not only in the more or less anxious safeguarding of the image that we want to have of ourselves, but also in a certain kind of relationship with the external world; in this case we are dealing with an object relation of a narcissistic type, namely, a link with situations, things or persons that serve as symbols to help reassure ourselves about our identity. Now, in narcissistic personality disorders, the feeling of identity is so precarious (the self is so scarcely cohesive, Kohut would say) that the patient finds it difficult to feel existent and is afraid of completely losing contact with himself or herself if deprived of such reassurances. These include what Kohut calls 'self-objects', namely, objects of a narcissistic type that are experienced as neither internal nor external with respect to the bounds of the identity of the person. The psychoanalyst writes about a patient (Mr. W.):

> It was at such times, when his unsupported childhood self began to feel frighteningly strange to him and began to crumble, that he had in fact surrounded himself with his possessions—sitting on the floor, looking at them, checking that they were there: his toys and his clothes. And he had at that time a particular drawer that contained his things, a drawer he thought about sometimes at night when he could not fall asleep, in order to reassure himself. (Kohut, 1977, pp. 167–168)

In a category of clinical cases less serious than full-blown narcissistic personality, the individual who suffers from an insufficient sense of identity, while not being forced to adopt a defensive style that can give rise to pathological problems, can lead a normal life only by placing himself within a situation of dependence, and hence by eschewing positions of responsibility. This is an indication that narcissistic problems, in attenuated forms, are ubiquitous, and thus rather than narcissistic personalities, we should address the more or less effective ways in which each of us comes to deal with the difficult problem of our narcissistic equilibria.

* * *

Let us consider now another personality disorder: the borderline personality disorder. This pathology is of great interest to us here since it has been described as a disorder of *attachment*, with symptoms of fear of abandonment, and of intense, unstable relationships; as a disorder of *self*, with symptoms of identity disturbance, chronic feelings of emptiness, and dissociative states under stress; and as a disorder of *self-regulation*, with symptoms of impulsivity, suicidal behavior, self-mutilation, affective instability and difficulty controlling anger.

As to etiopathogenesis, the presence of impairments in mentalization and autobiographical memory is a key to understanding the broad spectrum of dysfunctions related to the self, including disturbances in self-narratives, commonly observed in borderline patients.

Firstly, disruptions of early attachment experiences can derail the development of first-person mentalization. For the hypothesis has been made that the absence of empathic affect-regulative-mirroring interactions (see Sect. 4.2.2) may prevent children from creating the necessary mappings between the emerging causal representations of emotional states in others and emerging distinct emotional states in themselves; this may in turn

give rise to a compromised representational system for internal self states (see Hernik, Fearon, & Fonagy, 2009, p. 148).

Secondly, the presence of impairments in first-person mentalization may hamper the forming of a self-concept as an organized, coherent and unified autobiographical self-representation. Building on Erikson's work, Otto Kernberg observed that borderline patients often vacillate between extremely positive and negative representations of self and of others. In terms of the above discussed self-memory system, borderline patients may be able to cluster their episodic information in the slots of the autobiographical knowledge base, but they fail to integrate this information into a distinct self-concept, unifying both positive and negative aspects of the self. As a result, borderline patients exhibit 'a fragmentation of the narrative self' (Fuchs, 2007): a self-narrative that consists of different, mostly poorly elaborated *current* self-concepts, which are activated in turn, depending on the situation they are in, or on who they are with (see Masterson & Klein, 1989; Kernberg, Selzer, Koenigsberg, Carr, & Appelbaum, 1989; cit. in van den Broeck, 2014). The borderline patient's self-image or sense of self, therefore, is 'markedly and persistently unstable' (APA, 2013, p. 664), incoherent, and discontinuous. In this perspective, switching self-concepts is a strategy to preserve the present fragile structure, because it is a potentially adaptive response to self-discrepant information (see van den Broeck, 2014, p. 199).

* * *

Finally, it should be noted that a diminishing of the feeling of existing, to the point of a 'conversion of ourselves to nothingness' (James, 1950, p. 293), may be the result not only of a psychopathological process; an analogous outcome can arise from a sudden breakdown of self-esteem, or from unexpected emotional upheavals, or when the continuity of the tissue of our sociality is broken, as can happen when one is suddenly thrown into a dehumanizing 'total institution'.

Two classic books on this topic are *The Informed Heart* by Bruno Bettelheim (1960) and Erving Goffman's *Asylums* (1961). In *The Informed Heart*, Bettelheim describes his struggle to curb the menace of self-disintegration as a concentration camp prisoner in Germany during the late 1930s. *Asylums* is a work of ethnographic research at a mental hospital, offering an analysis of the pre-patient and inpatient phases

of the 'moral career' of the mental patient, consisting of 'a series of abasements, humiliations and profanations of self' (1961, p. 14). In such self-mortifying circumstances, a set of strategies can be implemented that are designed to restore a sense of autonomy and self-worth to the institutionalized individual: he strives to cling to his memories, to a sense of dignity, or to the secret security of an affiliation. However, if all of these fail him, then he realizes that his mind has become empty, and not only does he no longer know who he is, but he also literally loses the feeling of being present (see Jervis, 1997, p. 36).

5.3 Scaling Up: Culture as a System of Defense Techniques

To recapitulate, a person knows that she exists insofar as she knows that she exists in a certain way, as a describable identity, constant over changes. But self-consciousness as finding oneself again as a known identity, as a feeling of biographical continuity, is not a psychological faculty guaranteed once for all, but it is a precarious acquisition; it is, in Giddens' words, 'something that has to be routinely created and sustained in the reflexive activities of the individual' (1991, p. 52). This precariousness renders the defensiveness immanent to the selfing process intelligible.

The construction and defense of subjective identity, however, is not only a developmental and clinical theme; defense mechanisms fall along a spectrum that stretches from the individual to the collective level. In order to explore such an enlargement of perspective, we turn our attention to Ernesto de Martino's research at the intersection of philosophy and anthropology.

At the core of de Martino's thought lie the questions we have been dealing with thus far: the *precarious* nature of the subject's self-construction and of the resulting *defensive* character of self-consciousness. He forges a phenomenological psychology of identity hinged on the concepts of *presence* and (the complementary) *crisis of presence*. 'Presence' is the feeling of existing, that is, the primary feeling of the presence of the self to itself; or self-consciousness as finding oneself at the center of one's own orderly and meaningful subjective world, and hence at the center of a historical and cultural environment to which one feels one belongs. But this

self-consciousness as the full certainty on which the experience and order of everyday living rest is a precarious acquisition, continuously constructed by culture and constantly exposed to the risk of crisis, that is, 'the existential drama of the being-there exposed to the risk of not being there' (de Martino, 2007, p. 115).

In his 1948 ethnohistorical study *The Magic World*, de Martino engages in the project of re-founding ethnology as a historical science. But the fulfillment of such a project requires that a limitation of the traditional reflection on consciousness—that is, taking presence as a *given* (see Sect. 3.4)—be overcome. To this claim for givenness the philosopher objects that presence ('the person's unitary being' or, in Kantian terms, 'the transcendental unity of self-consciousness') includes in itself its opposite in the form of the risk of its disintegration:

> …even the supreme principle of the transcendental unity of self-consciousness involves a supreme risk to the person, that is, the risk of losing the supreme principle that constitutes and grounds it. This risk arises when the person, instead of retaining her autonomy in her relationship to the contents, abdicates the task and allows the contents to assert themselves, outside the synthesis, as undominated elements, as given facts in an absolute sense. (Ibid., pp. 158–159)

In a passage of the *Analytic of Concepts* Kant envisaged the possibility of the loss of presence (i.e., the 'original synthetic unity of apperception'). But this possibility was taken into account not as a real risk, but only as an absurd consequence of failing to recognize that unity—in such a case, Kant writes, 'I would have as multicolored, diverse a self as I have representations of which I am conscious' (1998, B134, pp. 247–248). However, de Martino comments on Kant's attitude about the possibility of the disintegration of the person's unitary being, arguing that 'the claim forming the backbone of the second chapter of *The Magic World* interprets as a real existential risk what in Kant's criticism is only controversial argument' (de Martino, 2008, p. 21).

In brief, Kant's apperception is regarded as an immediate, ahistorical datum. In contrast, de Martino aims to show the genesis of apperception, its historicity:

> Kant assumed as a uniform historical given the analytical unity of apperception—that is, the thought of the I that does not vary in its contents but comprehends them as its own, and he posited the transcendental condition of this given in the synthetic unity of apperception. But as elements and data of consciousness do not exist (except perhaps by abstraction), so there does not exist any presence, any empirical being-there, that might be a datum, an original immediacy beyond all risk and incapable within its own sphere of any sort of drama and of any development—that is, of a history. (De Martino, 2007, p. 159; transl. from Ginzburg, 1991, p. 45)

Presence, therefore, is not a datum but a task, 'the human task of being there'; and this requires that we go beyond 'what passes away by letting it pass in forms of cultural coherence' (de Martino, 1995, p. 101). Absorbed in the study of magic and religiosity, de Martino viewed these phenomena as part and parcel of a collective effort to create culture, which meant for him attributing shape and meaning to the flux of life, creating value out of what passes away despite or against us, transcending the material poverty of everyday life. In a word, presence is cultural dynamism, 'movement that transcends the situation in value' (ibid., p. 103).

However, if presence is movement, crisis is inactivity, a stagnation of the valorizing activity. The 'critical moments of becoming' are just those situations in which the inertia of presence, which is tantamount to its loss, becomes an imminent threat. This may occur in the confrontation with death, in cases of psychological dissociation, alienation or loss of subjectivity (see Saunders, 1993). In all such moments 'the risk of not being there is more intense, and therefore cultural redemption is more urgent' (de Martino, 1953–1954, pp. 18–19).

These moments of crisis are adumbrated by a total reaction that is the *anxiety* that 'underlines the threat of losing the distinction between subject and object, between thought and action, between representation and judgment, between vitality and morality: it is the cry of one who is wobbling on the edge of the abyss' (de Martino, 1956, p. 25; transl. from Saunders, 1995, p. 332). Anxiety is the condition of the individual who feels paralysis afflicting a presence that is not able to move beyond a particular situation; it is the dread of 'not being able to be there in the world, to give to itself

a culturally possible world, to emerge from the situation, to transcend it through value, to lose presence and world' (de Martino, 1995, p. 110). A fragment of that anxiety can at times lurk within the folds of everyday life, for example, in the bewilderment that each of us may feel upon waking. In this connection, de Martino (1964) quotes from the *Recherche*, when the narrating 'I' describes how he happened to wake at midnight not knowing where he was, or even who he was, lost in an existential abyss in which he felt 'more bereft than a cave-dweller'. But soon the crisis diminishes:

> …the memory—not yet of the place where I was, but of several of those where I had lived and where I might have been—would come to me for help from on high to pull me out of the void (*néant*) from which I could not have got out on my own; I passed over centuries of civilization in one second, and the image confusedly glimpsed of oil-lamps, then of wing-collar shirts, gradually recomposed my self's original features. (Proust, 2002, p. 9)

The path that Proust elegantly describes—from the giddiness of total disorientation to the recovery of himself and of the world—illustrates the *reverse* of the delusional experience of change that announces the psychotic event, for, in this case, the backdrop of domesticity is destructured against any effort of recovery: 'thus a painful inversion of sign is in the process of gaining the most obvious and familiar perceptive areas, which now appear to be strange, bizarre, artificial, theatrical, unreal, mechanical, out of joint, absurd' (de Martino, 1964, p. 143). And that inversion of sign reflects 'the fall of the presentificating energy on all the fronts of the possible valorisation' (ibid.).

In the psychopathological crisis, anxiety expresses the resistance that presence opposes to its annihilation, that is, to the regression into 'biological vitality' which, in opposition to culture, is chaos, confusion and madness (de Martino, 2002, p. 657). In experiencing the extreme risk of 'resubmerging themselves in nature, in the complete wreckage of the human', the patient attempts to exert control over such a risk by virtue of 'the de-historification of becoming—or more precisely, of what is happening as current or possible negativity' (de Martino, 2015, p. 103). In other words, he suspends becoming within himself, striving to carry out a total escape from the historicity of existence. Such de-historification can be noticed, for example, in the reaction of stupor:

> …a schizophrenic was realizing, with growing anxiety, that insurmountable difficulties thwarted his action: any movement that he was about to make seemed to present the perilous possibility of committing a harmful or ineffective act; and thus this mental patient, dominated by anxiety, chose not to eat, dress, or wash, finally reducing himself to the absolute immobility of catatonic stupor. (De Martino, 2008, pp. 32–33)

Yet this search for total absence is an unproductive strategy, an inadequate defense mechanism, to the extent that it is not able to carry out the redemption of presence, that is, to reintegrate it into the historical reality. Thus the psychopathological condition turns out to be a merely private 'individual drama' of escape from history, which is unable to 'reestablish the spiritual dialectic' (de Martino, 1956, p. 20). Psychotic patients then fail to 'retake possession of the alienated psychic realities, putting them once again into the cultural circuit, redisclosing to them their values' (ibid.; transl. by Saunders, 1995, p. 332).

In contrast with the psychopathological ('irrelative') de-historification is the de-historification that is put to use under cultural control. In order to resolve the critical moments of becoming only culture can offer 'an organic system of vital techniques of defense', which are all particular forms of the fundamental technique of 'institutional' de-historification, that is, the suspension of becoming in the pure iteration of myth and ritual:

> Magical protection […] is carried out thanks to the institution of a metahistorical level that absolves two distinct protective functions. Above all, this level creates a stable and traditionalized representative horizon in which the risky variety of possible individual crises finds a moment of coming to a halt, configuration, unification and cultural reintegration. At the same time, the metahistorical functions as a place of the 'de-historification' of becoming: a place in which, through the repetition of identical operative models, the historical proliferation of happening can from one time to the next be reabsorbed, and thus amputated of its actual and possible negativity. (De Martino, 2015, p. 94)

Let us take, for example, the reaction of bewilderment to a mournful event. When individuals face such a distressing reality, which is beyond their control, culture offers them a path along which their bereavement

is experienced, but at the same time overcome, that is, in which there is the crisis ('the crisis of grief') but also redemption from the crisis. This is the mythical-ritual de-historification: a descent into hell, but with the knowledge that one will escape it. An itinerary stretching into a world that is no longer the historical world (the world of everyday uncertainties and of great crises of existence) but rather the timeless world of myth (a body of scriptural and oral narratives and symbols) and its ritual repetition. A world in which there is death but also resurrection, and in which one pursues the narrative of death and resurrection because this narrative allows one to tell oneself that death can always be overcome.

In Sect. 5.1 we introduced de Martino's notion of territorial anguish, which can be regarded as the spatial or geographic parameter of feeling of being in crisis. The crisis of grief is the paradigmatic exemplification of a further parameter, the *temporal* one. In the loss of a beloved person there is a temporal fracture, that is, a traumatic breakdown of the expected continuity, which brings about an inner destructuration, the loss of ego boundaries. Like Laing's schizophrenic, the mourner has the sensation of losing herself (her own presence) due to the disruption of biographical continuity.[11] And similarly to autobiographical reasoning which is capable of compensating for threats of self-discontinuity in times of biographical change and rupture (see Sect. 4.3.2), death rituals serve to bring the mourners back into their particular history by assimilating the crisis of grief to a metahistorical pattern (the narrative of death and resurrection).

* * *

De Martino's anthropological work anticipated what we have seen of the centrality of identity in infant research, in social, personality and dynamic psychology, and in psychopathology. In this perspective, his phenomenological psychology of identity can be combined with the view of defense mechanisms that we have been delineating in this chapter.

As has been argued, the ultimate root of the primary defensiveness of the subject is the precariousness of self-consciousness as description of identity—the primary and universal existential risk of the loss of

[11] In the last years of his life, de Martino (2002) was concerned with the theme of the delusions of the end of the world (see Wetzel, 1922), a theme that summarizes the elements of territorial loss and loss of temporal continuity.

presence, as de Martino put it. Human self-conscious subjectivity constitutes itself as a repertoire of composite psychological maneuvers, of activities that take pains to cope with its lack of ontological guarantee, constructing itself on the edge of its original 'non-being', as it were. Within this framework, de Martino's anthropological inquiry into the social mechanisms (in particular those of ritual) that allow communities and individuals to defend themselves from anxiety at critical moments of becoming, joins forces with dynamic psychology to probe the uncertainties that concern self-consciousness: uncertainties regarding self-image, the acceptability of oneself 'as one is', one's inner solidity (see Jervis, 1997, p. 33).

In this perspective, the theme of presence and its crisis is a matter that is certainly historical but also, and perhaps still more, biological and psychological. Accordingly, it is no longer only the community that, living in history and making culture, forges the techniques to protect presence. Defense mechanisms fall along a spectrum that stretches from the individual level to the collective. The individual mind, far from being conceived as the place of an unproductive 'irrelative dehistorification', is now that cunning sphere of intrapsychic defenses and interpersonal maneuvers to which each of us appeals, in our relationships with other people and with our environment, in order to defend our own self-describability and, indissolubly, the solidity of our own self-conscious being. At a social and collective level, on the other hand, defenses consist in the construction of a system of references (in part symbolic and ritual) which give perspective to living, domesticity and meaning to one's own 'being-in-the-world'.

Now, all these individual and collective defensive structures, all these 'systems of presence', produce functional balances or adaptations, and any intra- or interpersonal psychological balance is precarious, fluctuating, modifiable and ultimately always deficient and unsatisfying for the individual. On this point, it is well worth noting, we have to turn once more to the lucid pessimism with regard to which Freud was so clear. Because he believed not in a definitive synthetic conciliation of the lacerations of the human psyche, but rather in the endless quest for increasingly better—yet never completely harmonious—balance.

5.4 A Robust Theory of the Self

In this chapter, we have investigated the idea of the human subject underlying the psychodynamic inquiry into defenses. The self as subjective identity is a construction with no metaphysical guarantee; it is not something guaranteed once and for all, but rather a precarious acquisition, continuously under construction by a human organism and constantly exposed to the risk of dissolution. This precariousness is the key to grasping the defensive nature of identity self-construction. The need to construct and protect an identity that is valid to the greatest extent possible is rooted in the primary need to subsist subjectively, and thus to exist solidly as a describable ego, as a unitary subject. And what we have seen is that identity self-construction is indeed the cornerstone of human development across the entire lifespan. Thus, far from being an epiphenomenon, a representation of one's inner life that plays no role in the intrinsic dynamics of the body, the incessant construction and reconstruction of an acceptable and adaptively functioning identity is the process that produces our intra- and interpersonal balances, and is thus the foundation of psychological well-being and mental health. Unlike the continuously self-rewriting autobiographies of the Joycean machine, identity as a story of the self is by no means contingent and evanescent; it is *a layer of personality that represents a causal center of gravity*.

Thus, the psychodynamic component of our theory affords the development of a 'robust' (i.e., genuinely realist) view of the self. In contrast to Dennett's deflationist conception of the self, according to which the self is a mere abstraction, analogous to a non-existent, but pragmatically useful, physical center of gravity, there is an open alternative, a realist or somewhat inflationist position compatible with everything Dennett, and eliminativists (or antirealists) in general, have to say about the architecture of the human neurocognitive system. There certainly exists something like Dennett's 'Joycean machine', that is, the subpersonal machinery that supports the construction and reconstruction of a narrative self, but this does not prevent us from claiming that the complex, persistent and highly structured psychological effects of this machinery are a *real* phenomenon, insofar as they constitute a causal center of gravity. In this sense, on our

view, it would be better to speak of a *robust* Joycean machine, thereby suggesting that the multiple, widely distributed and chaotic nature of the brain processes that constitute the Joycean machine does not entail that, at a higher level, the unity of the mind is simply an illusion.

Note that in positing such a 'robust' Joycean machine, we are exploiting the same kind of cognitive-science findings that Dennett invokes, together with many other data from developmental, dynamic, social and personality psychology, in order to build a realist theory of the self. Antirealists are unable to acknowledge the causal efficacy of the self for at least two reasons. First, they do not take into account the psychodynamic ingredient and the related teleology of the selfing process, which turns out to be much larger and much more important than they suppose. Indeed, while according to, for example, Dennett (2014), the self only serves to solve 'little problems of interpersonal activity', we have seen that identity self-construction is so important that it can be regarded as the keystone of the development of the whole existence of the individual. Second, eliminative or antirealist theories render the existence of a Cartesian Ego, whose characteristics are not matched by its neuronal counterpart, highly implausible, but have no impact on our naturalist, bottom-up and relational view of the self.

Our criticism of Dennett's eliminativism has much in common with that of Jenann Ismael. Ismael advances a model of the self that is midway between the Dennettian picture of mind/brain as a termite colony, and a naïve Cartesianism. What Ismael supports in Dennett's model is the substitution of the Cartesian theater by the Joycean machine. What she takes issue with is Dennett's claim that the autobiographical monologue assembled by the Joycean machine ('the Joycean stream') is just a fiction developed for the external audience. On the contrary, the Joycean self-centered streams of consciousness—the inner presentation of the subject's first personal representation of the world—play 'an important and substantial role in the intrinsic dynamics of the body' (Ismael, 2006, p. 348). Here Ismael builds on the difference between self-organizing and self-governing systems mentioned in Sect. 3.1.2, and her view of the evolution of the human cognitive architecture as emerging from 'a line of development that leads from simpler systems to self modelers' (ibid., p. 352) is fully consonant with many aspects of our bottom-up approach.

Scientific evidence, therefore, offers an at best shaky defense of eliminative narrativism. As noted earlier, the empirical data that the anti-realist exploits in constructing her theory simply *underdetermine* what one is able to say about the ontological status of the self. And things being so, one could hold that the brain produces a real narrative self that is not causally inert. A stronger reason to deny the causal efficacy of the self, taken as a psychobiological and psychosocial product, seems to come from metaphysics rather than from the sciences and is connected with a reductionist view of mental causation. The reductive strategy aims to a *reductio ad unum* of this complexity, and it may be based on two main premises: (1) that only bottom-level (neural) processes may be endowed with causal efficacy; and (2) that there is no chance to reduce the kind of multilevel and composite processes we appeal to in order to explain the workings of the self to bottom-level (neural) processes.

Taking for granted the truth of (2), the conclusion that the self is just an epiphenomenal entity rests on the first premise. Premise (1), however, is very contentious, as the debate on mental causation clearly shows. Here again, we are not addressing the metaphysical issue directly, but simply underlining the complexity of the debate involved.

A purely metaphysical approach should indeed settle many intricate issues—for example, specific controversial theses such as the causal closure of the physical domain thesis, and the no overdetermination thesis—and take on the challenge of general (and almost intractable) issues such as (1) the nature of causal *relata*, (2) the nature of properties, (3) the metaphysics of the nomological, (4) the nature of substances, (5) the nature of causal relation itself, and so on and so forth (see Gibb, 2013 for a much longer and more detailed list). We believe that hoping to arrive at a definitive conclusion with regard to such difficult problems—in isolation from the concrete and successful scientific explanations of behavior—is an overly optimistic expectation. This is why we find the adoption of a sort of 'negotiation model' more attractive: the construction of a metaphysical picture of the mental causation realm should always involve a continuous trade-off between metaphysical considerations on the one hand, and the suggestions offered by our successful explanatory practices on the other (see Di Francesco & Tomasetta, 2015). This strategy—which we adopted when dealing with the interface problem—should not

be endorsed in virtue of its conciliatory aspect, but rather because it is strongly suggested by the fact that purely metaphysical principles represent a far from a stable basis for our theoretical constructions. In any case, whatever one may think of this 'pragmatic' attitude, we would like to underline a crucial point: reference to science as it is actually performed (and not to science as it should be, according to prior—questionable—metaphysical insights) may be taken as the starting point of a general argument in favor of a comprehensive view of causal explanation, sensitive both to the metaphysical and the epistemological issues involved. And as we have shown, the contemporary science of the self is a multifaceted, pluralistic and many-level enterprise, and no monodimensional philosophical approach to it is a good candidate to offer a comprehensive theory of such a complex subject matter.

And so, at the end of our analysis of the mechanisms of self-conscious subjectivity, we find that we can reject the Cartesian conception of the subject but still propose a realist picture of the self. The self is a process, the objective, biological-cognitive process of reflexivity which emanates from the dialectic between the Jamesian I and Me. And the Me that the synthesizing I-process makes is not an epiphenomenon, but rather a layer of personality that serves as a causal center of gravity in the history of the agent.

6

Epilogue

In the last chapter, the presentation of our robust theory of the self came to an end. It is now time to take stock and try to locate our views within a wider context.

Throughout the entire book we have argued against the view of self-consciousness as a basic modality of consciousness, as a primary, elemental, simple awareness of the self, preceding any other form of knowing. Against this idealistic view, Sect. 3.4 suggested that self-consciousness is a knowing of being-there in a certain way, a self-describing, an identity forming. This is an integrative selfing process, a synthetic function. But it is not a Kantian synthetic function insofar as it invokes a psychological level of analysis that is not guaranteed by a transcendental level of analysis; thus, it involves the empirical subject which – as we maintained in Chap. 5 – is primarely non-unitary and pursues its unity in the act of mobilizing resources against the threat of disgregation. As de Martino puts it, the unity of apperception is a *task*, 'the human task of being-there'.

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8_6

The result of the synthesizing I-process is the Me-self, whose developmental story has been told within a multidimensional—naturalistic, bottom-up, and systemic-relational—framework.

The most minimal form of the Me is bodily self-awareness. This is a representation of one's own body, taken as a whole, which is analogical and imagistic in nature—a level of representation which is neither nonconceptual nor fully conceptual. In Sect. 4.3 we suggested that this nonverbal, analogical representation of the bodily self acts as a fixed referent around which autobiographical memories can start being organized. The Me to which the infant begins to attach episodic memories is the Jamesian material self. Self-narratives, therefore, do not create selves. The autobiographical self as a continuity across time and space that is interpreted reflectively by the agent would not arise without the material self.

On the other hand, it is the psychological unity—and notably the unity of an autobiographical narrative—that constitutes ourselves as Lockean persons, that is, as morally responsible agents. As seen in Sect. 4.3.1, there can be cases of dissociation among the Jamesian selves: agents may possess a material self and a social self but lack a spiritual self. In such cases, the agent considers herself responsible only insofar as she is held socially responsible for her actions. By contrast, she is never fully able to responsibly appropriate the products of her own mind, given their difficulty in constructing an introspective experiential space.

In Sect. 4.3.2, we saw that the diachronic dimension of introspective self-consciousness evolves with the development of a repertoire of social-cognitive competencies that enable the synthesizing I-process to take the form of autobiographical reasoning, the process through which the Me-self as a biography of the self is formed and used. Although a life narrative begins to emerge in middle childhood, the complexity and coherence of this narrative increase across adolescence.

This narrative of self-identity has an essential psychodynamic component. People's self-defining life stories have an intrinsically defensive nature; the description-narration of one's own inner life is organized on the basis of the fundamental need to construct and defend a self-image endowed with an at least minimal solidity. Thus, far from being an epiphenomenal,

transient phenomenon—a fictional character invented to facilitate predictions of behavior without any real correlate (Dennett's ephemeral virtual captain)—the incessant construction and reconstruction of an acceptable and adaptively functioning identity is the process that produces our intra- and inter-personal balances, and hence serves as a foundation of psychological well-being and mental health. The selfing process, therefore, imposes a teleology of self-defense on the human psychobiological system; it is the ongoing construction of a system of defenses, the continuously renovated capacity to curb and cope with anxiety and disorder.

Thus, in our psychodynamic perspective, self-identity construction is reminiscent of Kierkegaard's 'struggle of being against non-being' (cit. in Giddens, 1991, p. 48). This view allows us to dissent radically from the poststructural and/or postmodern rhetoric of the 'death of the subject'. This rhetoric arises from a maneuver which can be traced back to Romantic ideology, and consists in a dismantling of the classic credibility of the subject, and a subsequent suggestion of a multiplicity of the self, a decentralization of the ego, a polymorphism of identities and, in short, the end of a cohesive image of the mind. This project of weakening the ego or, still more radically, of shattering identity, characterizes a line of thought that, after incubating in the artistic avant-gardes, reached its climax in the Parisian culture of the 1970s. Its most radical formulation is found in Gilles Deleuze and Félix Guattari's *L'Anti-Œdipe*. These two thinkers go so far as to celebrate the fragmentation, multiplicity and discontinuity of the self in psychosis; on their view, schizophrenia is revolutionary, at both a social and individual level.

This intellectual operation, which today lingers only as the oddity of an age of folly and unlimited arrogance, would not be worth mentioning were it not that, since the 1990s, ideas not so distant from it have fostered the postmodernist and socio-constructivist reflection on identity (see Seigel, 2005). A well-known example is Gergen (1991), according to which the postmodern identity is multiple, shattered, bereft of any reality except for what is socially constructed moment to moment in everyday interactions. And in this view, it is all to the good; in fact, the multiplicity of the self (which he describes as the 'multiphrenic

condition') is to be accentuated in order to allow the subject to expand herself in different directions, to evolve and to create ever new opportunities of personal growth (see also Rose, 1996).

Nothing could be more in contrast with our conception of the human subject. If we accept the claim of the self-defensive nature of self-consciousness, any project of a weakening of the self or, worse, a disintegration of the identity, shows itself in its true light, that is, as an apology for mental suffering, as a failure to appreciate the tragic dimension of psychosis (see Glass, 1993; Jervis, 2011).

Our rejection of the eliminativists' and post-modernists' radical anti-realism of the self resulted from combining insights from narrative constructivism with the claim that the narrative self has causal efficacy. Yet our narrative approach still seems to be vulnerable to an antirealist objection. Such a narrativism, so the objection goes, is an approach that puts normative constraints on our self-narratives—constraints such as 'narrative unity' or 'narrative coherence'. But then, Kristjánsson notes, 'it is difficult to shake the suspicion that a person may possess a completely coherent self-identity that is nevertheless false' (2010, p. 39). Thus we are required to offer criteria by which *self-knowledge* may be distinguished from *self-deception*.

On the other hand, this distinction is required by the 'unmasking trend' mentioned in Sect. 2.2. When we use the cognitive sciences as a source of tools to set up a criticism of the subject, our guiding principle is that project of *demystification* (the systematic search for self-deception and the uncovering of underlying truth) which lies at the core of the critical tradition to which Freud belongs—that is, ultimately, the secular, rationalist, individualist culture of modernity. It is true, therefore, that Freud taught us that the description-narration of our inner life comes to be organized on the basis of a self-apologetic defensiveness, and hence is a construction permeated by myths and interested self-deceptions. Nevertheless, to this claim he always associated the firm belief that this self-image can be at least partially demystified, thus acknowledging the possibility of a path toward *genuine self-knowledge*. Thus, unlike those trends of thought that cultivate a radically conventionalist view of knowledge, the tradition to which Freud belongs draws a clear-cut line of demarcation between 'historical truth' and 'narrative truth'.

In this regard, the personological conception of narrative identity introduced in Sect. 5.2 can be helpful. According to McAdams, the narrative self is a third layer of personality built upon the dispositional base and motivational, social-cognitive and developmental adaptations. During personality development, internalized and evolving stories of the self layer over adaptations, which layer over traits; and this process of layering may be *integrative*: 'Traits capture the actor's dramaturgical present; goals and values project the agent into the future. An autobiographical author enters the developmental picture […] to integrate the reconstructed past with the experienced present and envisioned future' (McAdams, 2015, p. 226). Here the selfing process is a search for itself that strives for *a synthesis of the various strata of personality* which is reminiscent of the ancient ideal of *eudaimonia*: the good person as a healthy, fully functioning, self-realized person.

A criterion that affords a distinction of self-knowledge from self-deception is thus the following: deceptive self-narratives are those that fail to play an integrative role within the three-tiered structure of personality. In this perspective, telling a coherent self-story is not enough. A fully coherent but false self-narrative is a 'façade' marked by bad faith, something inauthentic and two-dimensional which tends to pass itself off (in accordance with our irrepressible tendencies to self-deception) as the 'solid', or 'deep', structure of the person. In McAdams' terms, such a self-narrative fails to integrate with the other layers of personality.

We may illustrate this perspective by returning to Locke's definition of responsibility as the capacity of *critically* re-appropriating one's own acts, projects, memories. In Sect. 4.1 the Lockean theory of inner sense was challenged in light of Carruthers' version of the self-other parity view; if the ISA theory is well grounded, we have a very strong constraint on the construction of a theory of moral responsibility congruent with the findings of cognitive sciences: the existence of conscious propositional attitudes, such as judgments and decisions, cannot be among the theory's commitments. Thus, to give only one example, let us consider the theories of the 'real self' (e.g., Frankfurt, 1988; Watson, 2004). These theories claim that an agent can be held responsible exclusively for those actions that have been motivated by psychological states that have been (or would be) endorsed, for it is such

endorsements that disclose our identity as practical agents. But as King and Carruthers (2012) observe, if the psychological states that define the agent's real self are conscious thoughts, their elimination implies the non-existence of the real self.

But now let us try to reconceptualize Frankfurt's discourse on practical identity in terms of our view of self-interpretation as a narrative re-appropriation of the products of the unconscious machinery. Within this framework—which gives a nonconsciousness-dependent direction to Locke's theory of person—the Lockean re-appropriation can be defined as 'critical' only in the sense of being a self-narrative that is more 'honest' (less imbued with 'bad faith') than what we usually practice. In other words, the critical, or rather responsible, re-appropriation of one's own actions and mentations (and more in general of one's own life events) consists in a process of self-knowledge that goes beyond (and against) the mechanisms of self-deception underlying self-conscious subjectivity. On the other hand, this is already implicit in Frankfurt's perspective, where individuals establish a *critical engagement* with their psychological lives, and the appropriation of the real self is a path of self-knowledge.

This idea can be more clearly stated by returning to the contrast between guilt and responsibility set up in Sect. 4.2.2. It has often been noticed in psychoanalysis that a person who suffers from a sense of guilt deceives herself by treating what she feels guilty about as extraneous to herself; in short, she expels it from her self-narrative. Let us return to the driver who, after running over the poor pedestrian, is afflicted by a tormenting sense of guilt, and longs for absolution (see Sect. 4.2.2). In his feeling guilty he represents that event to himself as a foreign body, perceives it as a discontinuity in the flux of his life—in the psychoanalytic idiom, he 'evacuates' it. By contrast, if that individual admits the fact that, say, he is a person whose overbearing and aggressive character reverberates in his way of driving, as well as the fact that when he ran over the pedestrian he was driving too fast (until that time he was 'culpably ignorant' of these facts), by so doing he takes a path toward a responsible appropriation of the fatal event that dispels its egodystonic character. And thus, whereas the sense of guilt is the outcome of a self-narrative permeated by bad faith, the assumption of responsibility is the result of a path of self-knowledge that finally permits him to include in his own life

story (i.e., his narrative identity) also the crimes or misdemeanors that he has committed. And this path of self-knowledge does entail the subject's 'reflective endorsement', which is reconceptualized, however, in terms of an autobiographical reasoning in which illusions and self-deceptions are rooted out and dispelled.

Thus, in the end, in contesting the prerogatives of the Lockean consciousness we do not reduce the responsibility of a person who would no longer be her own master. What is lost with the decentering of the subject can be regained by means of a 'psychotherapeutic' self-narrative that puts the individual before an 'inner court'. Thus, Locke's critical re-appropriation becomes a form of demystifying hermeneutics which has its measure of objectivity in a dynamic psychology informed by cognitive sciences:

> Over against illusion and the fable-making function, demystifying hermeneutics sets up the rude discipline of necessity. It is the lesson of Spinoza: one first finds himself a slave, he understands his slavery, he rediscovers himself free within understood necessity (Ricoeur, 1970 [1965], p. 35).

References

Addis, D. R., & Tippett, L. J. (2008). The contributions of autobiographical memory to the content and continuity of identity. In F. Sani (Ed.), *Self-continuity: Individual and collective perspectives* (pp. 71–84). New York: Psychology Press.

Allen, C., & Trestman, M. (2015). Animal consciousness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2015/entries/consciousness-animal/

Al-Namlah, A. S., Fernyhough, C., & Meins, E. (2006). Sociocultural influences on the development of verbal mediation: Private speech and phonological recoding in Saudi Arabian and British samples. *Developmental Psychology, 42*, 117–131.

Antaki, V. C. (Ed.) (1981). *The psychology of ordinary explanations in social behaviour*. London: Academic Press.

Antaki, V. C. (1985). Ordinary explanation in conversation: Causal structures and their defence. *European Journal of Social Psychology, 15*, 213–230.

APA (American Psychiatric Association). (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Armstrong, D. (1968). *A materialist theory of the mind*. London: Routledge.

Astuti, R. (2001). Are we natural dualists? A cognitive developmental approach. *Journal of the Royal Anthropological Institute, 7*(3), 429–447.

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8

Averill, J. R. (1980). A constructivist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience: Vol. 1. Theories of emotion* (pp. 305–339). New York: Academic Press.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Baars, B. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences, 6*, 47–52.

Baars, B. (2003). How brain reveals mind: Neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies, 10*, 100–114.

Bacon, F. (2000). *The New Organon* (L. Jardine, Ed.). Cambridge: Cambridge University Press. (orig. ed. 1620).

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*, 159–186.

Balint, M. (1965). Early developmental states of the ego. Primary object love. In *Primary love and psycho-analytic technique*. London: Tavistock, pp. 90–108. (orig. ed. 1937).

Balint, M. (1992). *The basic fault. Therapeutic aspects of regression*. Evanston, IL: Northwestern University Press. (orig. ed. 1968).

Barrett, J. L. (2004). *Why would anyone believe in god?* Plymouth: AltaMira Press.

Bauer, P. (2014). The development of forgetting: Childhood amnesia. In P. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (pp. 519–544). West Sussex, UK: Wiley.

Bechtel, W., Abrahamsen, A., & Graham, G. (1998). The life of cognitive science. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 1–104). Oxford: Blackwell.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 1–62). New York: Academic Press.

Bermúdez, J. L. (1995). Transcendental arguments and psychology. *Metaphilosophy, 26*, 379–401.

Bermúdez, J. L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.

Bermúdez, J. L. (2001). Nonconceptual self-consciousness and cognitive science. *Synthese, 129*(1), 129–149.

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. London: Routledge.

Bermúdez, J. L. (2007). Self-consciousness. In M. Velmans & S. Schneider (Eds.), *The blackwell companion to consciousness* (pp. 456–467). Oxford: Blackwell.

Bermúdez, J. L. (2009). Self: Body awareness and self-awareness. In W. P. Banks (Ed.), *Encyclopedia of consciousness* (Vol. 2, pp. 289–300). Oxford: Elsevier.

Bermúdez, J. L. (2011). Bodily awareness and self-consciousness. In S. Gallagher (Ed.), *Oxford handbook of the self* (pp. 157–179). Oxford: Oxford University Press.

Bettelheim, B. (1960). *The informed heart: Autonomy in a mass age*. Glencoe, IL: The Free Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bloom, P. (2004). *Descartes' baby*. New York: Basic Books.

Bodei, R. (2002). *Destini personali*. Milan: Feltrinelli.

Bottenberg, F. (2012). Review of "the self and its emotions". *Philosophical Psychology, 26*(3), 480–484.

Bowlby, J. (1969/1982). *Attachment and loss: Vol. 1. Attachment*. New York: Basic Books.

Bowlby, J. (1973). *Attachment and loss: Vol. 2. Separation: Anxiety and anger*. New York: Basic Books.

Bowlby, J. (1980). *Attachment and loss: Vol. 3. Loss: Sadness and depression*. London: Hogarth Press and Institute of Psycho-Analysis.

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton and Oxford: Princeton University Press.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies, 61*, 127–148.

Boyer, P. (2001). *Religion explained. The evolutionary origins of religious thought*. New York: Basic Books.

Brandon, P. (2014). Body and self: An entangled narrative. *Phenomenology and the cognitive sciences*, forthcoming.

Brentano, F. (1995). *Psychology from an empirical standpoint* (2nd ed.). London: Routledge. (orig. ed. 1874).

Bretherton, I., & Munholland, K. A. (2008). Internal working models in attachment relationships: Elaborating a central construct in attachment theory. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research and clinical applications* (pp. 102–130). New York: Guilford Press.

Brownell, C. A., Svetlova, M., & Nichols, S. R. (2012). Emergence and early development of the body image. In C. Brownell & V. Slaughter (Eds.), *Early development of body representation* (pp. 37–58). Cambridge: Cambridge University Press.

Brüne, M. (2005). 'Theory of mind' in schizophrenia: A review of the literature. *Schizophrenia Bulletin, 31*, 21–42.

Buckner, C., Shriver, A., Crowley, S., & Allen, C. (2009). How "weak" mind-readers inherited the earth. *Behavioral and Brain Sciences, 32*, 140–141.

Carpendale, J., & Lewis, C. (2006). *How children develop social understanding*. Oxford: Blackwell.

Carruthers, P. (2009a). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*, 121–138.

Carruthers, P. (2009b). Mindreading underlies metacognition. *Behavioral and Brain Sciences, 32*, 164–176.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Carruthers, P. (2013a). Mindreading in infancy. *Mind & Language, 28*(2), 141–172.

Carruthers, P. (2013b). Mindreading the self. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds* (3rd ed., pp. 467–485). Oxford: Oxford University Press.

Carruthers, P. (2014). The fragmentation of reasoning. In P. Quintanilla, C. Mantilla, & P. Cépeda (Eds.), *Cognición Social y Lenguaje: La intersubjetividad en la evolución de la especie y en el desarrollo del niño*. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Carruthers, P. (2016). Who's in charge anyway? Published on the *OUP blog* on 08/01/15 at: http://blog.oup.com/2015/08/whos-in-charge-conscious-mind/

Carruthers, P. (2017). The illusion of conscious thought. In D. Jacquette (Ed.), *The Bloomsbury companion to the philosophy of consciousness*. London: Bloomsbury Press.

Carruthers, P., Fletcher, L., & Ritchie, B. (2012). The evolution of self-knowledge. *Philosophical Topics, 15*, 13–37.

Carruthers, P., & Ritchie, B. (2012). The emergence of metacognition: Affect and uncertainty in animals. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition* (pp. 76–93). Oxford: Oxford University Press.

Carver, C. S. (2012). Self-awareness. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 50–68). New York-London: The Guilford Press.

Cassam, Q. (1997). *Self and world*. Oxford: Clarendon Press.

Castel, R. (1973). *Le psychanalisme*. Paris: Maspero.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 27–42.

Cohen, E., & Barrett, J. L. (2008). When minds migrate. Conceptualizing spirit possession. *Journal of Cognition and Culture, 8*(1–2), 23–48.

Cole, M., Gay, J., Glick, J. A., & Sharp, D. W. (1971). *The cultural context of learning and thinking*. New York: Basic Books.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language, 53*(4), 594–628.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107*(2), 261–288.

Conway, M. A., Singer, J. A., & Tagini, A. (2004). The self and autobiographical memory: Correspondence and coherence. *Social Cognition, 22*, 491–529.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology, 64*, 201–229.

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., & Smith, J. D. (2009). Metacognition is prior (Commentary on Carruthers). *Behavioral and Brain Sciences, 32*, 142.

Courage, M. L., Edison, S. C., & Howe, M. L. (2004). Variability in the early development of visual self-recognition. *Infant Behavior and Development, 27*, 509–532.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Crawford, T. N., Cohen, P., Johnson, J. G., Sneed, J. R., & Brook, J. S. (2004). The course and psychosocial correlates of personality disorder symptoms in adolescence: Erikson's developmental theory revisited. *Journal of Youth and Adolescence, 33*, 373–387.

Cummins, R. (1997). *Representations, targets and attitudes*. Cambridge, MA: MIT Press.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Damasio, A. (2010). *Self comes to mind*. New York: Pantheon.

Davidson, D. (1982). Paradoxes of irrationality. In R. Wollheim & J. Hopkins (Eds.), *Philosophical essays on Freud* (pp. 289–305). Cambridge: Cambridge University Press.

Davidson, D. (1985). Deception and division. In J. Elster (Ed.), *The multiple self* (pp. 79–92). Cambridge: Cambridge University Press.

De Martino, E. (1951–1952). Angoscia territoriale e riscatto culturale nel mito achilpa delle origini. Contributo allo studio della mitologia degli aranda. *Studi e Materiali di Storia delle religioni*, *23*, 51–66.

De Martino, E. (1953–1954). Fenomenologia religiosa e storicismo assoluto. *Studi e materiali di storia delle religioni*, 24–25, 1–25.
De Martino, E. (1956). Crisi della presenza e reintegrazione religiosa. *Aut Aut, 31*, 17–38.
De Martino, E. (1964). Apocalissi culturali e apocalissi psicopatologiche. *Nuovi Argomenti, 69–71*, 105–141.
De Martino, E. (1995). *Storia e metastoria*. Lecce: Argo.
De Martino, E. (2002). *La fine del mondo* (2nd ed.). Turin: Einaudi.
De Martino, E. (2005). *The Land of Remorse* (D. L. Zinn Trans.). London: Free Association Books. (orig. ed. 1961).
De Martino, E. (2007). *II mondo magico*. Turin: Bollati Boringhieri. (orig. ed. 1948).
De Martino E. (2008). *Morte e pianto rituale*. Turin: Bollati Boringhieri. (orig. ed. 1958).
De Martino, E. (2015). *Magic. A theory from the South* (D. L. Zinn Trans.). Chicago: The University of Chicago Press. (orig. ed. 1959).
De Villiers, J. (2013). Language and reasoning about beliefs. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world* (pp. 96–100). New York: Oxford University Press.
DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge: Cambridge University Press.
Dehaene, S. (2014). *Consciousness and the brain*. New York: Viking.
Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*, 200–227.
Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*, 204–211.
Delgado, J. M. (1969). *Physical control of the mind*. New York: Harper and Row.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown.
Dennett, D. C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). Hillsdale, NJ: Erlbaum.
Dennett, D. C. (2005). *Sweet dreams*. Cambridge, MA: MIT Press.
Dennett, D. C. (2009). Intentional systems theory. In B. P. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford: Oxford University Press.

Dennett, D. C. (2016). Artifactual selves: A response to Lynn Rudder Baker. *Phenomenology and Cognitive Sciences, 15*(1), 17–20.

Dennett, D. C., & Akins, K. (2008). Multiple drafts model. *Scholarpedia, 3*(4), 4321.

Dennett, D. C., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences, 15*, 183–247.

Descartes, R. (1984). Author's replies to the first set of objections. (J. Cottingham, D. Murdoch, & R. Stootho, Eds.). *The philosophical writings of Descartes* (Vol. 2, pp. 74–86). Cambridge: Cambridge University Press. (orig. ed. 1641).

De Vignemont, F. (2007). Habeas corpus: The sense of ownership of one's own body. *Mind & Language, 22*, 427–449.

De Villiers, J. (2013). Language and reasoning about beliefs. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the Social World* (pp. 96–100). New York: Oxford University Press.

Di Francesco, M. (2007). Extended cognition and the unity of mind: Why we are not 'spread into the world'. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind* (pp. 213–227). Berlin: Springer.

Di Francesco, M., & Marraffa, M. (2013). Consciousness, the unconscious, and the ego illusion. *Dialogues in Philosophy, Mental and Neuro Sciences, 6*(1), 10–22.

Di Francesco, M., & Marraffa, M. (2014). A plea for a more dialectical relationship between personal and subpersonal levels of analysis. *Frontiers in Psychology, 5*, 1165.

Di Francesco, M., & Piredda, G. (2012). *La mente estesa*. Milan: Mondadori Università.

Di Francesco, M., & Tomasetta, A. (2015). The end of the world? Mental causation, explanation and metaphysics. *Humana.Mente, 29*, 167–190.

Di Francesco, M., Marraffa, M., & Paternoster, A. (2014). Real selves? Subjectivity and the subpersonal mind. *Phenomenology and Mind, 7*, 118–133.

Dodds, E. (2004). *The Greeks and the irrational*. Berkeley-Los Angeles, CA: University of California Press. (orig. ed. 1951).

Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Dunphy-Lelii, S., & Wellman, H. (2012). Delayed self-recognition in autism: A unique difficulty? *Research in Autism Spectrum Disorders, 6*(1), 212–223.

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness*. New York: Academic Press.

Eagle, M. N. (2003). The postmodern turn in psychoanalysis: A critique. *Psychoanalytic Psychology, 20*(3), 411–424.

Eagle, M. N. (2011). *From classical to contemporary psychoanalysis: A critique and integration*. New York: Routledge.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Chichester, UK: Wiley and Sons.

Ellenberger, H. F. (1970). *The discovery of the unconscious: The history and evolution of dynamic psychiatry*. New York: Basic Books.

Engelbert, M., & Carruthers, P. (2010). Introspection. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 245–253.

Erikson, E. H. (1968). *Identity: Youth and crisis*. New York: Norton.

Farmer, H., & Tsakiris, M. (2012). The bodily social self: A link between phenomenal and narrative selfhood. *Review of Philosophy and Psychology, 3*(1), 125–144.

Federn, P. (1952). *Ego psychology and the psychoses*. New York: Basic Books. (orig. ed. 1927).

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General, 113*, 464–486.

Fernyhough, C. (2009). What can we say about the inner experience of the young child? (Commentary on Carruthers). *Behavioral and Brain Sciences, 32*, 143–144.

Fernyhough, C., Bland, K. A., Meins, E., & Coltheart, M. (2007). Imaginary companions and young children's responses to ambiguous auditory stimuli: Implications for typical and atypical development. *Journal of Child Psychology and Psychiatry, 48*, 1094–1101.

Fivush, R. (2010). The development of autobiographical memory. *Annual Review of Psychology, 62*(2), 2–24.

Fodor, J. A. (1991, June). Too hard for our kind of mind? *London Review of Books, 27*, 12.

Fodor, J. A. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Fogelin, R. J. (1985). *Hume's skepticism in the treatise on human nature*. London: Routledge.

Fogelin, R. J. (2009). *Hume's skeptical crisis. A textual study*. Oxford: Oxford University Press.

Fonagy, P., Gergely, G., & Target, M. (2007). The parent-infant dyad and the construction of the subjective self. *Journal of Child Psychology and Psychiatry, 48*, 288–328.

Fonagy, P., Gergely, G., Jurist, E., & Target, M. (2002). *Affect regulation, mentalization, and the development of the self*. London: Other Press.

Frankfurt, H. (1988). *The importance of what we care about*. Cambridge: Cambridge University Press.
Frechette, G. (2013). Searching for the self: Early phenomenological accounts of self-consciousness from Lotze to Scheler. *International Journal of Philosophical Studies, 21*(5), 654–679.
Freud, S. (1914). On narcissism: An introduction. In Strachey (1957), Vol. 14, pp. 67–102.
Freud, S. (1915). Instincts and their vicissitudes. In Strachey (1956–1974), Vol. 14, pp. 117–40.
Freud, S. (1923). The ego and the id. In Strachey (1956–1974), Vol. 19, pp. 12–66.
Freud, S. (1929/1930). Civilization and its discontents. In Strachey (1956–1974), Vol. 21, pp. 57–146.
Freud, S. (1933). New introductory lectures on psycho-analysis. In Strachey (1964), Vol. 22, pp. 1–267.
Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and Cognition, 21*(1), 52–54.
Frith, C., Blakemore, S.-J., & Wolpert, D. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews, 31*, 357–363.
Fuchs, T. (2007). Fragmented selves: Temporality and identity in borderline personality disorder. *Psychopathology, 40*(6), 379–387.
Gallagher, S. (2005). *How the body shapes the mind*. London: Oxford University Press.
Gallagher, S. (Ed.) (2011). *The Oxford handbook of the self*. Oxford: Oxford University Press.
Gallagher, N., & Meltzoff, A. (1996). The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology, 9*, 211–233.
Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. London: Routledge.
Gallagher, S., & Zahavi, D. (2015). Phenomenological approaches to self-consciousness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2015/entries/self-consciousness-phenomenological/
Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*, 396–403.
Gallino, L. (2006). *Dizionario di sociologia*. Turin: Utet.

Gardner, S. (1999). Psychoanalysis, contemporary views. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 683–685). Cambridge, MA: MIT Press.

Gardner, S. (2000). Psychoanalysis and the personal/sub-personal distinction. *Philosophical Explorations, 3*, 96–119.

Gennaro, R. (2012). *The consciousness paradox*. Cambridge, MA: MIT Press.

Gergely, G. (2002). The development of understanding self and agency. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 26–46). Oxford: Blackwell.

Gergely, G. (2004). The social construction of the subjective self: The role of affect-mirroring, markedness, and ostensive communication in self development. In L. Mayes, P. Fonagy, & M. Target (Eds.), *Developmental science and psychoanalysis* (pp. 45–82). London: Karnac.

Gergely, G., & Unoka, Z. (2008a). Attachment, affect-regulation and mentalization. In E. L. Jurist, A. Slade, & S. Bergner (Eds.), *Mind to mind: Infant research, neuroscience and psychoanalysis* (pp. 50–88). New York: Other Press.

Gergely, G., & Unoka, Z. (2008b). The development of the unreflective self. In F. N. Busch (Ed.), *Mentalization. Theoretical considerations, research findings and clinical implications* (pp. 57–102). New York-London: The Analytic Press, Taylor and Francis Group.

Gergely, G., & Watson, J. S. (1996). The social biofeedback theory of parental affect-mirroring: The development of emotional self-awareness and self-control in infancy. *International Journal of Psychoanalysis, 77*(6), 1181–1212.

Gergely, G., & Watson, J. S. (1999). Early social-emotional development: Contingency perception and the social biofeedback model. In P. Rochat (Ed.), *Early social cognition* (pp. 101–137). Hillsdale, NJ: Erlbaum.

Gergely, G., Koós, O., & Watson, J. S. (2010). Contingent parental reactivity in early socio-emotional development. In T. Fuchs, H. C. Sattel, & P. Henningsen (Eds.), *The embodied self: Dimensions, coherence and disorders* (pp. 141–169). Stuttgart: Schattauer.

Gergen, K. (1991). *The saturated self*. New York: Basic Books.

Gibb, S. G. (2013). Introduction. In S. G. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental causation and ontology* (pp. 1–17). Oxford: Oxford University Press.

Giddens, A. (1991). *Modernity and self-identity*. Cambridge: Polity Press.

Giddens, A. (1992). *The transformation of intimacy*. Cambridge: Polity Press.

Gill, M. M., & Holzman, P. S. (Eds.) (1976). *Psychology versus metapsychology: Essays in memory of George S. Klein*. New York: International University Press.

Ginzburg, C. (1991). Momigliano and de Martino. *History and Theory, 30*(4), 37–48.

Glass, J. M. (1993). *Shattered selves: Multiple personality in a postmodern world*. Ithaca, NJ: Cornell University Press.

Goffman, E. (1959). *The presentation of self in everyday life*. Edinburgh: University of Edinburgh.

Goffman, E. (1961). *Asylums. Essays on the social situation of mental patients and other inmates*. New York: Doubleday.

Goffman, E. (1963). *Behavior in public places*. New York: The Free Press.

Goldberg, A. (1984). The tension between realism and relativism in psychoanalysis. *Psychoanalysis and Contemporary Thought, 7*, 367–386.

Goldman, A. I. (2006). *Simulating minds*. Oxford: Oxford University Press.

Gopnik, A. (1993). How we read our own minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.

Greenwald, A. G. (1992). New look 3: Unconscious cognition reclaimed. *American Psychologist, 47*, 766–779.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories*. Chicago, IL: University of Chicago Press.

Griffiths, P. (2004a). Instinct in the '50s: The British reception of Konrad Lorenz's theory of instinctive behaviour. *Biology and Philosophy, 19*(4), 609–631.

Griffiths, P. (2004b). Is emotions a natural kind? In R. C. Solomon (Ed.), *Thinking about feeling: Contemporary philosophers on emotions* (pp. 233–249). Oxford: Oxford University Press.

Griffiths, P. E., & Scarantino, A. (2009). Emotions in the wild: The situated perspective on emotion. In P. Robbins & M. Aydede (Eds.), *Cambridge handbook of situated cognition* (pp. 437–453). Cambridge: Cambridge University Press.

Grünbaum, A. (1984). *The foundations of psychoanalysis*. Berkeley: University of California Press.

Habermas, T. (2011). Autobiographical reasoning: Arguing and narrating from a biographical perspective. *New Directions for Child and Adolescent Development, 131*, 1–17.

Habermas, T., & Köber, C. (2015a). Autobiographical reasoning is constitutive for narrative identity: The role of the life story for personal continuity. In K. C. McLean & M. Syed (Eds.), *The Oxford handbook of identity development* (pp. 149–165). Oxford: Oxford University Press.

Habermas, T., & Köber, C. (2015b). Autobiographical reasoning in life narratives buffers the effect of biographical disruptions on the sense of self-continuity. *Memory, 23*(5), 664–674.

Habermas, T., & de Silveira, C. (2008). The development of global coherence in life narratives across adolescence: Temporal, causal, and thematic aspects. *Developmental Psychology, 44*, 707–721.

Hamlin, J. K. (2013a). The origins of human morality: Complex sociomoral evaluations by preverbal infants. In J. Decety & Y. Christen (Eds.), *Research and perspectives in neurosciences* (pp. 165–188). Berlin: Springer.

Hamlin, J. K. (2013b). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science, 22*(3), 186–193.

Harré, R. (1986). Mind as social formation. In J. Margolis, M. Krausz, & R. M. Burien (Eds.), *Rationality, relativism and the human sciences* (pp. 91–106). Dordrecht: Nijhoff.

Harré, R. (1987). The social construction of selves. In K. Yardley & T. Honess (Eds.), *Self and identity: Psychological perspectives* (pp. 41–52). New York: Wiley.

Haugeland, J. (1998). *Having thought*. Cambridge, MA: MIT Press.

Hernik, M., Fearon, P., & Fonagy, P. (2009). There must be more to development of mindreading and metacognition than passing false belief tasks. *Behavioral and Brain Sciences, 32*, 147–148.

Hoerl, C. (2007). Episodic memory, autobiographical memory, narrative: On three key notions in current approaches to memory development. *Philosophical Psychology, 20*, 621–640.

Holt, R. R. (1989). *Freud reappraised. A fresh look at psychoanalytic theory*. New York: Guilford.

Hopkins, J. (1988). Epistemology and depth psychology: Critical notes on 'the foundations of psychoanalysis'. In P. Clark & C. Wright (Eds.), *Mind, psychoanalysis and science* (pp. 33–60). Oxford: Blackwell.

Howe, M. L. (2011). *The nature of early memory: An adaptive theory of the genesis and development of memory*. Oxford-New York: Oxford University Press.

Howe, M. L. (2014). The co-emergence of the self and autobiographical memory. In P. J. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (pp. 545–567). Wiley-Blackwell: Hoboken, NJ.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin, 113*, 305–326.

Howe, M. L., & Courage, M. L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104*, 499–523.

Howe, M. L., Courage, M. L., & Rooksby, M. (2009). The genesis and development of autobiographical memory. In M. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (pp. 177–196). Hove, UK: Psychology Press.

Hull, C. L. (1943). *The principles of behavior*. New York: Appleton-Century-Crofts.

Hume, D. (2000). *A treatise of human nature* (D. F. Norton & M. J. Norton, Eds.). Oxford: Oxford University Press. (orig. ed. 1739–1740).

Hutto, D. D. (2008). *Folk psychological narratives*. Cambridge, MA: The MIT Press.

Ismael, J. T. (2006). Saving the baby: Dennett on autobiography, agency, and the self. *Philosophical Psychology, 19*, 345–360.

Ismael, J. T. (2011). Self-organization and self-governance. *Philosophy of the Social Sciences, 41*(3), 327–351.

James, W. (1950). *The principles of psychology*. New York: Dover. (orig. ed. 1890).

Jervis, G. (1969). Contributo allo studio dell'isteria. Psicopatologia della crisi di possessione. *Il lavoro neuropsichiatrico, 37*(3), 555–572.

Jervis, G. (1984). *Presenza e identità*. Milan: Garzanti.

Jervis, G. (1989). *La psicoanalisi come esercizio critico*. Milan: Garzanti.

Jervis, G. (1993). *Fondamenti di psicologia dinamica*. Milan: Feltrinelli.

Jervis, G. (1996). Pregiudizio. In *Enciclopedia delle scienze sociali, Vol. 6* (pp. 771–776). Treccani: Rome.

Jervis, G. (1997). *La conquista dell'identità*. Milan: Feltrinelli.

Jervis, G. (2001). *Psicologia dinamica*. Bologna: il Mulino.

Jervis, G. (2002). Un commento ai due scritti di Morris Eagle e di Robert Holt. *Rassegna di Psicologia, 19*(2), 45–51.

Jervis, G. (2006). Identità. In A. Zamperini, F. Barale, V. Gallese, S. Mistura, & M. Bertani (Eds.), *Psiche. Dizionario storico di psicologia, psichiatria, psicoanalisi e neuroscienze* (pp. 504–509). Turin: Einaudi.

Jervis, G. (2007). The unconscious. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind* (pp. 147–158). Berlin: Springer.

Jervis, G. (2011). In G. Corbellini & M. Marraffa (Eds.), *Il mito dell'interiorità*. Turin: Bollati Boringhieri.

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences, 364*, 2325–2335.

Jung, C. G. (1971). Psychological types (G. Adler & R. F. C. Hull, Eds.). *The collected works of C. G. Jung* (Vol. 6). Princeton: Princeton University Press. (orig. ed. 1928).

Kant, I. (1998). *Critique of pure reason*. Cambridge: Cambridge University Press. (orig. ed. 1781–1787).

Kapitan, T. (1999). The ubiquity of self-awareness. *Grazer Philosophische Studien, 57*, 17–44.

Kernberg, O. F. (1984). *Severe personality disorders*. New Haven, CT: Yale University Press.

Kernberg, O., Selzer, M. A., Koenigsberg, H. W., Carr, A. C., & Appelbaum, A. H. (1989). *Psychodynamic psychotherapy of borderline patients*. New York: Basic Books.

Kim, S., Fonagy, P., Allen, J., Martinez, S., Iyengar, U., & Strathearn, L. (2014). Mothers who are securely attached in pregnancy show more attuned infant mirroring 7 months postpartum. *Infant Behavior and Development, 37*(4), 491–504.

King, M., & Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy, 9*, 200–228.

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. *Progress in Brain Research, 164*, 257–264.

Kitcher, P. (2011). *Kant's thinker*. Oxford: Oxford University Press.

Klein, G. S. (1976). *Psychoanalytic theories: An exploration of essentials*. New York: International Universities Press.

Kobak, R., Zajac, K.,& Madsen, S. (2016). Attachment disruptions, reparative processes, and psychopathology: theoretical and clinical Implications. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment* (pp. 25–39). New York-London: The Guilford Press.

Köber, C., Schmiedek, F., & Habermas, T. (2015). Characterizing lifespan development of three aspects of coherence in life narratives: A cohort-sequential study. *Developmental Psychology, 51*, 260–275.

Kohut, H. (1977). *The restoration of the self*. New York: International Universities Press.

Køster, A. (2016). Narrative and embodiment – a scalar approach. *Phenomenology and the Cognitive Sciences*, doi:10.1007/s11097-016-9485-8.

Kouider, S., Stahlhut, C., Gelskov, S., Barbosa, L., Dutat, M., de Gardelle, V., Christophe, A., Dehaene, S., & Dehaene-Lambertz, G. (2013). A neural marker of perceptual consciousness in infants. *Science, 340*, 376–380.

Kriegel, U. (2007). The phenomenologically manifest. *Phenomenology and the Cognitive Sciences, 6*(1–2), 115–136.

Kriegel, U. (2008). *The sources of intentionality*. Oxford: Oxford University Press.

Kristjánsson, K. (2010). *The self and its emotions*. Cambridge: Cambridge University Press.

Lackner, J. R., & Garrett, M. (1973). Resolving ambiguity: Effect of biasing context in the unattended ear. *Cognition, 1*, 359–372.

Lagattuta, K. H. (2014). Linking past, present, and future: Children's ability to connect mental states and emotions across time. *Child Development Perspectives, 8*(2), 90–95.

Laing, R. D. (1960). *The divided self: An existential study in sanity and madness*. London: Tavistock.

Laplanche, J., & Pontalis, J.-B. (1973). *The language of psycho-analysis*. New York: W.W. Norton and Co. (orig. ed. 1967).

Lewis, M. (1994). Myself and me. In S. Parker, R. Mitchell, & M. Boccia (Eds.), *Self-awareness in animals and humans: Developmental perspectives* (pp. 20–34). Cambridge: Cambridge University Press.

Lewis, M., & Brooks-Gunn, J. (1979). *Social cognition and the acquisition of the self*. New York and London: Plenum Press.

Lewis, M., & Carmody, D. P. (2008). Self-representation and brain development. *Developmental Psychology, 44*, 1329–1334.

Lewis, M., & Ramsay, D. (2004). Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child Development, 75*, 1821–1831.

Lichtenberg, J. D. (1989). *Psychoanalysis and motivation*. Hillsdale, NJ: Analytic Press.

Lichtenstein, H. (1977). *The dilemma of human identity*. New York: Jason Aronson.

Lind, S. (2010). Memory and the self in autism. A review and theoretical framework. *Autism, 14*(5), 430–456.

Livingstone Smith, D. (1999). *Freud's philosophy of the unconscious*. Dordrecht: Kluwer.

Locke, J. (1975). *An essay concerning human understanding*. Oxford: Clarendon Press. (orig. ed. 1694).

Loftus, E. F., & Ketcham, K. (1994). *The myth of repressed memory*. New York: St. Martin's Press.

Luria, A. (1976). *Cognitive development: Its cultural and social foundations*. Cambridge, MA: Harvard University Press. (orig. ed. 1974).

Lyyra, P. (2009). Two senses for 'givenness of consciousness'. *Phenomenology and Cognitive Sciences, 8*, 67–87.

MacIntyre, A. (1984). *After virtue*. Notre Dame: University of Notre Dame Press.

Mackenzie, C. (2008). Introduction. In K. Atkins & C. Mackenzie (Eds.), *Practical identity and narrative agency* (pp. 1–28). New York: Routledge.
Macmillan, M. (1997). *Freud evaluated: The completed arc*. Cambridge, MA: MIT Press.
Manson, N. C. (2000). 'A tumbling-ground for whimsies'? The history and contemporary role of the conscious/unconscious contrast. In T. Crane & S. Patterson (Eds.), *The history of the mind-body problem* (pp. 148–168). London: Routledge.
Marcia, J. E. (1966). Development and validation of ego-identity status. *Journal of Personality and Social Psychology, 5*, 551–558.
Marraffa, M. (2011a). Theory of mind. In J. Fieser (Ed.), *The internet encyclopedia of philosophy*. http://www.iep.utm.edu/theomind/
Marraffa, M. (2011b). Precariousness and bad faith. Jervis on the illusions of self-conscious subjectivity. *Iris, 3*(6), 171–187.
Marraffa, M. (2012). Remnants of psychoanalysis. Rethinking the psychodynamic approach to self-deception. *Humana.Mente, 20*, 223–243.
Marraffa, M. (2013). De Martino, Jervis, and the self-defensive nature of self-consciousness. *Paradigmi, 31*(2), 109–124.
Marraffa, M. (2014). The unconscious, self-consciousness, and responsibility. *Rivista internazionale di filosofia e psicologia, 5*, 207–220.
Marraffa, M. (2015). Mindreading and introspection. *Rivista internazionale di filosofia e psicologia, 6*, 249–260.
Marraffa, M., & Meini, C. (2016). *L'identità personale*. Rome: Carocci.
Marraffa, M., & Paternoster, A. (2013). *Sentirsi esistere*. Rome: Laterza.
Marraffa, M., & Paternoster, A. (2016). Disentangling the self. A naturalistic approach to narrative self-construction. *New Ideas in Psychology, 40*, 115–122.
Masterson, J., & Klein, R. (1989). *Psychotherapy of the disorders of the self: The Masterson approach*. New York: Brunner/Mazel.
Maynard-Smith, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.
McAdams, D. P. (1985). *Power, intimacy, and the life story: Personological inquiries into identity*. Homewood, IL: Dorsey Press.
McAdams, D. P. (1996). Personality, modernity, and the storied self: A contemporary framework for studying persons. *Psychological Inquiry, 7*, 295–321.
McAdams, D. P. (1997). The case for unity in the (post)modern self: A modest proposal. In R. D. Ashmore & L. Jussim (Eds.), *Self and identity. Fundamental issues* (pp. 46–78). New York: Oxford University Press.
McAdams, D. P. (2013). The psychological self as actor, agent, and author. *Perspectives on Psychological Science, 8*(3), 272–295.

McAdams, D. P. (2015). Tracing three lines of personality development. *Research in Human Development, 12*, 224–228.

McAdams, D. P., & Cox, K. S. (2010). Self and identity across the life span. In R. M. Lerner (Ed.), *The handbook of life-span development* (Vol. 2, pp. 158–207). New York: Wiley.

McAdams, D. P., & Olson, B. D. (2010). Personality development: Continuity and change over the life course. *Annual Review of Psychology, 61*, 517–542.

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist, 61*, 204–217.

McDowell, J. (1996). *Mind and world* (2nd ed.). Cambridge, MA: Harvard University Press.

McLean, K. C., & Syed, M. (2015). The field of identity development needs an identity. In Idd (Ed.), *The Oxford handbook of identity development* (pp. 1–10). Oxford-New York: Oxford University Press.

Mead, G. H. (1934). *Mind, self, and society*. Chicago: University of Chicago Press.

Meini, C. (2015). From cradle to internet. The social nature of personal identity. *Rivista internazionale di filosofia e di psicologia, 6*(2), 282–296.

Meins, E. (2011). Social relationships and children's understanding of mind: Attachment, internal states, and mind-mindedness. In M. Siegal & L. Surian (Eds.), *Access to language and cognitive development* (pp. 23–43). Oxford-New York: Oxford University Press.

Meltzoff, A. N., & Moore, M. K. (1995). Infants' understanding of people and things: From body imitation to folk psychology. In J. Bermudez, A. J. Marcel, & N. Eilan (Eds.), *The body and the self* (pp. 43–69). Cambridge, MA: MIT Press.

Merleau-Ponty, M. (2010). *Child psychology and pedagogy. The Sorbonne lectures 1949–1952*. Evaston, IL: Northwestern University Press.

Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Metzinger, T. (2011). The no-self alternative. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 279–296). Oxford: Oxford University Press.

Meyer, S., & Shore, C. (2001). Children's understanding of dreams as mental states. *Dreaming, 11*, 179–194.

Mitchell, S. A. (1988). *Relational concepts in psychoanalysis*. Cambridge, MA: Harvard University Press.

Moscovici, S. (2007). *Psychoanalysis: Its image and its public*. Cambridge: Polity Press. (orig. ed. 1961).

Musholt, K. (2013). Self-consciousness and nonconceptual content. *Philosophical Studies, 163*, 649–672.
Nelson, K. (1989). *Narratives from the crib*. Cambridge, MA: Harvard University Press.
Nelson, K. (2007). *Young minds in social worlds*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Oxford University Press.
Nielsen, M., Dissanayake, C., & Kashima, Y. (2003). A longitudinal investigation of self-other discrimination and the emergence of mirror self-recognition. *Infant Behavior and Development, 26*, 213–226.
Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology, 35*, 613–624.
Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice-Hall.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.
Oyama, S. (2000a). *The ontogeny of information: Developmental systems and evolution*. Durham, NC: Duke University Press. (orig. ed. 1985).
Oyama, S. (2000b). *Evolution's eye: A systems view of the biology-culture divide*. Durham, NC: Duke University Press.
Oyama, S., Griffiths, P. E., & Gray, R. D. (Eds.) (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge, MA: MIT Press.
Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford University Press.
Parker, S., Mitchell, R., & Boccia, M. (Eds.) (1994). *Self-awareness in animals and humans: Developmental perspectives*. Cambridge: Cambridge University Press.
Paternoster, A. (2010). Le teorie simulative della comprensione e l'idea di cognizione incarnata. *Sistemi Intelligenti, 22*(1), 131–161.
Paternoster, A. (2013). Un problema dell'inconscio cognitivo. *Sistemi Intelligenti, 25*(3), 469–483.
Pears, D. (1982). Motivated irrationality, Freudian theory and cognitive dissonance. In R. Wollheim & J. Hopkins (Eds.), *Philosophical essays on Freud* (pp. 279–288). Cambridge: Cambridge University Press.

Pereboom, D. (2014). Kant's transcendental arguments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/fall2014/entries/kant-transcendental/

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J., Zauner, P., & Sprung, M. (2005). What does 'that' have to do with point of view? Conflicting desires and 'want' in German. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 220–244). Oxford: Oxford University Press.

Peterfreund, E. (1978). Some critical comments on psychoanalytic conceptualizations of infancy. *The International Journal of Psychoanalysis, 59*, 427–441.

Piaget, J. (1929). *The child's conception of the world*. London: Routledge and Kegan Paul. (orig. ed. 1926).

Povinelli, D. J. (1995). The unduplicated self. In P. Rochat (Ed.), *The self in infancy: Theory and research* (pp. 161–192). Amsterdam: North-Holland/Elsevier Science Publishers.

Povinelli, D. J. (2001). The self: Elevated in consciousness and extended in time. In C. Moore & K. Lemmon (Eds.), *The self in time: Developmental perspectives* (pp. 75–95). Mahwah, NJ: Erlbaum.

Prebble, S. C., Addis, D. R., & Tippett, L. J. (2013). Autobiographical memory and sense of self. *Psychological Bulletin, 139*, 815–840.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.

Proust, M. (2002). *In search of lost time vol. 1: The way by Swann's*. London: Penguin Classics. (orig. ed. 1913).

Raffard, S., D'Argembeau, A., Lardi, C., Bayard, S., Boulenger, J. P., & Van der Linden, M. (2010). Narrative identity in schizophrenia. *Consciousness and Cognition, 19*, 328–340.

Reese, E., Yan, C., Jack, F., & Hayne, H. (2010). Emerging identities: Narrative and self from early childhood to early adolescence. In K. C. McLean & M. Pasupathi (Eds.), *Narrative development in adolescence: Creating the storied self* (pp. 23–44). New York: Springer.

Ricoeur, P. (1970). *Freud and philosophy: An essay on interpretation*. New Haven: Yale University Press. (orig. ed. 1965).

Ricoeur, P. (1994). *Oneself as another*. Chicago: University of Chicago Press.

Robbins, P. (2006). The ins and outs of introspection. *Philosophy Compass, 1*(6), 617–630.

Rochat, P. (2012). Primordial sense of an embodied self-unity. In V. Slaughter & C. Brownell (Eds.), *Early development of body representations* (pp. 3–18). New York: Cambridge University Press.

Rohlf, M. (2014). Immanuel Kant. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2014/entries/kant/

Rose, N. (1996). *Inventing our selves: Psychology, power, and personhood*. Cambridge: Cambridge University Press.

Rosenthal, D. M. (1990). On being accessible to consciousness. *Behavioral and Brain Sciences, 13*, 621–622.

Rossi, P. (1968). *Francis Bacon: From magic to science*. London: Routledge.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Ryle, G. (2009). *The concept of mind*. London: Routledge. (orig. ed. 1949).

Saunders, G. R. (1993). 'Critical ethnocentrism' and the ethnology of Ernesto De Martino. *American Anthropologist, 95*(4), 875–893.

Saunders, G. R. (1995). The crisis of presence in Italian pentecostal conversion. *American Ethnologist, 22*(2), 324–340.

Scarantino, A. (2012). Discrete emotions: From folk psychology to causal mechanisms. In P. Zachar & R. Ellis (Eds.), *Categorical and Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell* (pp. 135-154). Amsterdam: John Benjamins.

Schear, J. C. (2009). Experience and self-consciousness. *Philosophical Studies, 144*, 95–105.

Schechtman, M. (1996). *The constitution of selves*. Ithaca: Cornell University Press.

Schechtman, M. (2011). The narrative self. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 394–416). Oxford: Oxford University Press.

Schechtman, M. (2013). Identity, personal (philosophy of). In B. Kaldis (Ed.), *Encyclopedia of philosophy and the social sciences* (pp. 453–454). London: Sage.

Schlenker, B. R. (2012). Self-presentation. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 542–570). New York-London: The Guilford Press.

Schmidt-Hellerau, C. (2005). We are driven. *The Psychoanalytic Quarterly, 74*(4), 989–1028.

Schneider, S. (2007). Daniel Dennett on the nature of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 313–324). Oxford: Blackwell.

Schopenhauer, A. (1969). *The world as will and representation*. New York: Dover. (orig. ed. 1819).

Schroer, J. W., & Schroer, R. (2014). Getting the story right: A reductionist narrative account of personal identity. *Philosophical Studies, 171*, 445–469.

Schwalbe, M. L. (1993). Goffman against postmodernism: Emotion and the reality of the self. *Symbolic Interaction, 16*(4), 333–350.
Schwitzgebel, E. (2014). Introspection. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2014/entries/introspection/
Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard University Press.
Searle, J. (1990). Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences, 13*(4), 585–596.
Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Seigel, J. (2005). *The idea of the self: Thought and experience in Western Europe since the seventeenth century*. New York: Cambridge University Press.
Shani, I. (2007). Consciousness and the first person. *Journal of Consciousness Studies, 14*, 57–91.
Shoemaker, D. (2016). Personal identity and ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2016/entries/identity-ethics/
Sperber, D., & Hirschfeld, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences, 8*(1), 40–46.
Sprong, M., Schothorst, P., Vos, E., Hox, J., & van Engeland, H. (2007). Theory of mind in schizophrenia: Meta-analysis. *British Journal of Psychiatry, 191*, 5–13.
Sroufe, L. A. (1996). *Emotional development: The organization of emotional life in the early years*. Cambridge: Cambridge University Press.
Sternberg, R. J. (2012). Intelligence. *WIREs Cognitive Science, 3*, 501–511.
Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge, MA: MIT Press.
Stich, S. (1990). *The fragmentation of reason. Preface to a pragmatic theory of cognitive evaluation*. Cambridge, MA: MIT Press.
Strachey, J. (Ed.) (1956–1974). *The standard edition of the complete psychological works of Sigmund Freud*. London: Hogarth Press.
Strawson, P. F. (1959). *Individuals*. London: Methuen.
Strawson, P. F. (1966). *The bounds of sense*. London: Methuen.
Stroud, B. (1977). *Hume*. London: Routledge.
Suddendorf, T., & Butler, D. L. (2013). The nature of visual self-recognition. *Trends in Cognitive Sciences, 17*(3), 121–127.
Sullivan, H. S. (1953). *The interpersonal theory of psychiatry*. New York: Norton.
Tajfel, H., & Turner, J. C. (1986). The social identity theory of inter-group behavior. In S. Worchel & L. W. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago: Nelson-Hall.

Taylor, C. (1989). *Sources of the self: The making of the modern identity*. Cambridge, MA: Harvard University Press.

Tilly, C. (2006). *Why?* Princeton, NJ: Princeton University Press.

Tolman, E. (1948). Cognitive maps in rats and men. *The Psychological Review, 55*(4), 189–208.

Tomasetta, A. (2015). *Persone umane*. Rome: Carocci.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology, 11*, 375–424.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1), 1–25.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). Oxford: Oxford University Press.

Tye, M. (2003). *Consciousness and persons*. Cambridge, MA: MIT Press.

Tylor, E. B. (2010). *Primitive culture*. Cambridge: Cambridge University Press. (orig. ed. 1871).

Van den Broeck, K. (2014). *Specificity and vantage perspective of autobiographical memories in borderline pathology*. PhD dissertation, University of Leuven.

Van Gulick, R. (1995). Why the connection argument doesn't work. *Philosophy and Phenomenological Research, 55*, 201–207.

Vandekerckhove, M. M. (2009). Memory, autonoetic consciousness and the self: Consciousness as a continuum of stages. *Self and Identity, 8*(1), 4–23.

Voltolini, A. (2002). Why it is hard to naturalize attitude aboutness. In W. Hinzen & H. Rott (Eds.), *Belief and meaning. Essays at the interface* (pp. 157–179). Hänsel-Hohenhausen: Frankfurt.

Watson, G. (2004). *Agency and answerability*. Oxford: Oxford University Press.

Weary, G., & Arkin, R. (1981a). Attributional self-presentation. In J. H. Harvey (Ed.), *New directions in attribution research*. Hillsdale, NJ: Erlbaum.

Weary, G., & Arkin, R. (1981b). Attributional self-presentation. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 223–246). Hillsdale, NJ: Erlbaum.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: Harvard University Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of the will. *American Psychologist, 54*, 480–491.

Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology, 86*(6), 838–848.
Wells, G., & Petty, R. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology, 1*, 219–230.
Welsh, T. (2006). Do neonates display innate self-awareness? Why neonatal imitation fails to provide sufficient grounds for innate self and other awareness. *Philosophical Psychology, 19*, 221–238.
Welsh, T. (2007). Primal experience in Merleau-Ponty's philosophy and psychology. *Radical Psychology, 6*(1), http://www.radpsynet.org/journal/vol6-1/index.html
Welsh, T. (2013). *The child as natural phenomenologist. Primal and primary experience in Merleau-Ponty's psychology*. Evanston, IL: Northwestern University Press.
Wetzel, A. (1922). Das Weltuntergangserlebnis in der Schizophrenie. *Z f. d. g. Neurologie und Psychiatrie, 28*, 403–428.
Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784.
Wheeler, M. A., Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin, 121*(3), 331–354.
Williams, D. (2010). Theory of own mind in autism. Evidence of a specific deficit in self-awareness? *Autism, 14*(5), 474–494.
Williams, D., & Happé, F. (2010). Representing intentions in self and other: Studies of autism and typical development. *Developmental Science, 13*, 307–319.
Wilshire, B. (1969). Protophenomenology in the psychology of William James. *Transactions of the Charles S. Peirce Society, 5*(1), 25–43.
Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.
Wollheim, R. (1993). Desire, belief and Professor Grünbaum's Freud. In Id. (Ed.), *The mind and its depths* (pp. 91–111). Cambridge, MA: Harvard University Press.
Wright Mills, C. (1940). Situated actions and vocabularies of motive. *American Sociological Review, 5*, 904–913.
Zahavi, D. (2003). Phenomenology of self. In T. Kircher & A. S. David (Eds.), *The self in neuroscience and psychiatry* (pp. 56–75). Cambridge: Cambridge University Press.
Zahavi, D. (2005). *Subjectivity and selfhood*. Cambridge, MA: MIT Press.

Zahavi, D. (2007). Self and other: The limits of narrative understanding. In D. D. Hutto (Ed.), *Narrative and understanding persons* (pp. 179–201). Cambridge: Cambridge University Press.

Zahavi, D. (2009). Naturalized phenomenology. In S. Gallagher & D. Schmicking (Eds.), *Handbook of phenomenology and cognitive science* (pp. 2–19). Berlin: Springer.

Zahavi, D. (2012). The time of the self. *Grazer Philosophische Studien, 84*, 143–159.

Zahavi, D. (2014). *Self and other. Exploring subjectivity, empathy, and shame*. Oxford: Oxford University Press.

Zahavi, D. (2015). Self and other: From pure ego to co-constituted we. *Continental Philosophy Review, 48*(2), 143–160.

Index

Di Francesco, Marraffa, & Paternoster, *The Self and its Defenses*,
DOI 10.1057/978-1-137-57385-8

T