Remnants of Psychoanalysis. Rethinking the Psychodynamic Approach to Self-Deception*

Massimo Marraffa †
marraffa@uniroma3.it

ABSTRACT

This article reflects on the phenomenon of self-deception in the context of the psychodynamic approach to defense mechanisms. Building on Giovanni Jervis' criticism of psychoanalysis, I pursue the project of a full integration of that approach in the neurocognitive sciences. In this framework, the theme of self-deception becomes a vantage point from which to sketch out a philosophical anthropology congruent with the ontology of neurocognitive sciences.

1. Debunking the Unconscious

According to the Cartesian doctrine of the perfect transparency of the mind, the latter is simply *res cogitans*, and thought, its defining attribute, is explicated in terms of awareness (*conscientia*). As he writes in the *Replies to the second set of objections*: «I use the term 'thought' to include everything that is within us in such a way that we are immediately aware [*conscious*] of it» (Descartes, 1641/1988, p. 113). Here there is no margin for the notion of unconscious mentality: «there can be nothing within me of which I am not in some way aware» (1641/1988, p. 77).[1] Many philosophers will follow him.

During the second half of the 19th century, however, the unconscious insistently claims its own rights. Neurologists and psychiatrists had drawn

* This essay is one of a series of papers (see Marraffa, 2011a,b,c, forthcoming) in which I have been trying to reconstruct and develop Giovanni Jervis' work on three themes: the unconscious, consciousness, and identity.

† University Roma Tre, Italy.

[1] Interpretations of Descartes's account of consciousness (like everything else in his philosophy) differ significantly. Here we are following John Cottingham's authoritative interpretation (see, e.g., Cottingham, 1988, p. 153).

attention on 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 was shaping the early experimental psychology. It is comprehensible, then, that philosophers, psychologists and neuroscientists were bewildered about phenomena that appeared to be mental but went beyond the sphere of awareness and conscious control.

As Livingstone Smith (1999) has convincingly shown, during the second half of the 19th Century two strategies were adopted to reconcile the existence of supposed unconscious mental phenomena with the consciousness-dependent conception of mind. 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, namely «the total possible consciousness may be split into parts which coexist but mutually ignore each other» (James, 1890/1950, p. 206). The second option consisted in denying that such phenomena were genuinely *mental*; the evidence for the existence of unconscious mental states was reconceptualised as evidence for *neurophysiological dispositions* for genuinely (i.e., conscious) mental states.

The two strategies are still options in current Anglo-American philosophy. John Searle has recast the dispositionalist approach to unconscious mental states, whereas the "partitionist" approach to self-deception has revived the dissociationist option.[2] Let us focus on the latter.

Self-deception is traditionally viewed as a temporary impairment of *normal* belief-forming processes.[3] In addition, it is seen as a phenomenon that gives rise to two paradoxes: the "static" paradox and the "dynamic" one (see Mele, 1997). The partitionist approach to self-deception aims to dispel the static paradox by dividing the agent into two (or more) sub-agents, whose minds include the belief that *p* and the belief that non-*p* respectively. And it tries to

[2] See Livingstone Smith (1999, chapters 14–16). At p. 141 the author interestingly notes that Searle's idea that "the ontology of the unconscious is strictly the ontology of a neurophysiology capable of generating the conscious" coincides with what the physiologist Ewald Hering had claimed in 1870.

[3] «Normal», that is, «from the analytic philosopher's point of view, where the central important epistemic goal seems to be the generation of true beliefs» (Sage, forthcoming).

dissipate the dynamic paradox by postulating that the deceived sub-agent cannot access the deceiving sub-agent's activities.

Donald Davidson is often considered the main "partitioner", but actually his partitionism is very moderate. Davidson thinks that when one runs across such (apparent) absurdities of reason as akrasia or self-deception, the personal psychology framework is not to be given up in favor of the subpersonal one, but rather it must be enlarged or extended so that one can find somewhere else the rationality set out by the principle of charity. On this perspective, the division of the mind is a *metaphoric* device to coherently describe (within the personal-level explanatory framework) a phenomenon (self-deception) that otherwise would be unintelligible. As Davidson puts it, a mental division is nothing but "a metaphorical wall" that keeps two contradictory beliefs separate. Consequently, we do not need to postulate «two minds each somehow able to act like an independent agent»; it is sufficient to imagine «a single mind not wholly integrated; a brain suffering from a perhaps temporary self-inflicted lobotomy» (Davidson, 1998, p. 8).

A stronger version of partitionism – appropriately defined as "homuncularist" by Johnston (1988, p. 63) – was suggested by David Pears. Here the psychological partitioning is no longer Davidson's metaphorical wall; rather it is a conceptual reconstruction of Freud's second topographical model of the mind. The psyche is divided into a "main system" and a "sub-system"; the latter is «built around the nucleus of the wish for the irrational belief» and it is «organized like a person» (Pears, 1984, p. 87). Now, as Jon Elster points out, Pears ascribes to the sub-system an internal rationality («it is an efficient, quasi-altruistic manipulator of the main system» (Elster, 1984, p. 1388). And this implies that the sub-system both has all sorts of propositional attitudes regarding the main system, and it is «able to weigh and choose between alternative ways of satisfying the wishes of the main system» (*ibid.*). But then, Elster very properly concludes, «these requirements almost inexorably imply that the subsystem must have some kind of consciousness» (*ibid.*).[4]

Thus we find again here that same need of reabsorbing the discourse on the unconscious into the discourse on consciousness that led some fin-de-siècle researchers to reinterpret the evidence for unconscious mental states as evidence for the possibility of a *dédoublement* of consciousness. On the basis

[4] In this connection, see the entry "Topique" in Laplanche & Pontalis (1967), where it is rightly pointed out that Freud's second topographical model of the mind has an anthropomorphic character.

of such a conclusion, it might appear strange that Davidson's (1982) and Pears' (1982) theories of self-deception are offered as defenses of Freud's theory. For is it not true that Freud put forward a subpersonal psychology (a "metapsychology") that aimed to go beyond the psychology of consciousness? As a matter of fact, the psychological partitioning approach really captures an aspect of Freud's theory of the unconscious; but unfortunately, it is an aspect that – as we will now see – is the main limit of Freud's theory.

2. Troubles with the Freudian Unconscious

When, in the last decade of the 19th Century, Freud intervenes in the dispute on the unconscious, he takes sides against the predominant "consciousness-centric" mentalism and in favor of the reality of occurrent and intrinsically unconscious mental events. The Freudian theory of the unconscious is, therefore, *programmatically* against the psychological partitioning insofar as this treatment of self-deception remains – as we have argued – within an introspective-intuitive psychology of consciousness. The problem is that, *as a matter of fact*, Freud failed to get himself out of that psychology.

Freud's view of the relationship between conscious and unconscious mind is the ground of the conception of consciousness still dominant in the current non-specialized (and sometimes philosophical) culture. The common culture about the mind is a largely psychodynamic culture. Of course, this culture represents an advance on the Cartesian thesis of the transparency of the mind, which informs the image of human beings typical of 19th Century middle class ethics, against which Freud polemicized. According to Victorian anthropology the essence of the human being in its highest expression, that of "the civilized gentleman", lies in the full control exerted by self-consciousness over mind and behavior. But if this anthropology was dominated by the idea of consciousness (and conscious agency) so that a person could say «If I did it, it is *evidently* because I chose it, because I wanted to do it», in the average culture of the mind one realizes that people are tossed about by instances which they do not always control very well, so that sometimes 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 (Jervis, 2011, p. xxi).

The psychodynamic culture of the mind, therefore, makes an important correction to the idea of a psyche consisting in conscious and self-transparent

intentions; but it is only a partial correction. In the average culture of the mind, influenced by psychoanalytic psychodynamics, holds what was the most evident limitation of the Freudian view of the unconscious: the definition of the unconscious is still given "by difference" from – and in some respects also depending on – the definition of consciousness; the latter is taken as a self-evident, primary quality of the mind, although it is then criticized and "downsized" in comparison with the traditional idealistic conception. Accordingly, the Freudian mind «is still dominated by the model of the conscious elaboration of choices, and within it the unconscious plays its tricks here and there, but nothing more» (Jervis 2011, p. xxii). Like all the psychoanalytic ideas, the Freudian unconscious is a sort of enlargement or extension of the everyday commonsense psychological framework, which is a psychology of consciousness.[5]

(One might remark that in recent years a number of philosophers, influenced by Davidson, have argued that the extension of our ordinary psychological conception of mind is a strength of the psychoanalytic theory.[6] This move is the basis of a defence of psychoanalysis against well-known epistemological challenges.[7] But as will become clear in the next section, the metaphilosophy inspiring this essay rejects any form of antinaturalism that deprives science of the domain of the mental construed as a space of reasons rather than causes. In our perspective, the right question to ask is how and to what extent the folk-psychological conceptual framework should be rectified in light of neurocognitive sciences, in which – pace Kandel (2005) – not much of psychoanalytic theory can be integrated.)

[5] See Manson, who rightly notices 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» (2000, p. 163). And see also O'Brien and Jureidini, who argue that «[j]ust as much as the mental entities that parade across our consciousness, those that inhabit the [psychoanalytic] unconscious are [...] "personal-level" phenomena [...] in terms of their contents at least, unconscious ideas are conjectured to be indistinguishable from their conscious counterparts in all things save the fact that consciousness of them is absent» (2003, p. 143).

[6] These philosophers think that «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).

[7] In this perspective, «[p]sychoanalytic explanations, like ordinary psychological explanations, may be exempt from the epistemological and methodological standards of experimental science» (Manson, 2003, p. 179).

Furthermore, it should be emphasized that if Freud still preserves the primacy of consciousness, this is not because he develops a phenomenology, which has this consciousness as a methodological source of its investigation of reality. In other words, Freud does not develop a theory of subjectivity at all, and not 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 with a very original and sometimes informally imaginative idiom, places, forces and events in the Freudian mind (ego, id, super-ego, censorship, libido, cathexis, and so on) 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. But these objects are more metaphorical than real, more imagined than described. It might be said that the Freudian psyche is a collection of imaginary interfaces of the nervous system; his theory of mind is the psychologization of a very personal speculative-introspective neurology. During the development of his thought after 1900, the way in which the psychological dimension becomes autonomous from the neurological one – from which Freud had started – never becomes detached from an objectivistic (and indeed one could say: subjectively objectivistic) way of conceiving the mind (see Jervis, 2011, pp. xxii-xxiii).

Freud then claims to describe in accordance with a positivistic objectivism neurobiological mechanisms as constitutive of the mind. But although these mechanisms aim to explain many dimensions of the affective and emotional life, they are not supposed to explain consciousness. In spite of the unconscious and its energy-driven instincts, the Freudian adult self-consciousness is once more "assumed" or "given". So we find in his work 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.

Briefly, psychoanalysis is a personal psychology that is masked as subpersonal psychology.[8]

[8] This is the gist of the famous objection that Sartre makes to Freud, when he rejects the idea of a censor mechanism (see Sartre, 1943, pp. 87–88). If Sartre's criticism is translated into the idiom of the explanatory levels, we obtain the claim that psychoanalysis (and, we add, the homuncularist partitionism) moves from the personal level to the sub-personal one, «but it ends up having to re-

3. Consciousness as Seen from the Bottom Up

Today the response of a psychologist to the above-discussed discontents over psychoanalysis would be claiming that cognitive science can count on a genuinely subpersonal level of analysis – the information-processing level, wedged between the personal sphere of phenomenology and the subpersonal one of neural facts – which no longer takes consciousness as an unquestionable assumption, as a non-negotiable given fact. The cognitivist mind is a process of construction and transformation of *representations*; and a mental representation is an explanatory hypothesis in a computational theory of cognition; it is a structure of information (somehow encoded in the brain), which is individuated exclusively in terms of intra-theoretical functional criteria.[9] Cognitive scientists introduce mental representations to explain intelligent behavior not differently from what physicists do when they posit entities like spin, charm and charge.

Cognitive science, therefore, challenges the traditional link between consciousness and intentionality, thus opening a conceptual space to build a consciousness-independent conception of the unconscious. As Dennett (1991) puts it, first the cognitive scientists develop a theory of intentionality that is independent of and more fundamental than consciousness – a theory that treats equally any form of unconscious representational mentality; and then, they proceed to work out a theory of consciousness on that foundation. In this perspective, consciousness is «an advanced or derived mental phenomenon» and not, as Descartes wanted, «the foundation of all mentality» (Dennett, 1993, p. 193).

In viewing consciousness no longer as something that explains, but rather as something that needs to be explained, analyzed, dismantled, cognitive science amends the Freudian thought on the basis of Darwinian naturalism. Differently from Freud's introspective-intuitive description of the unconscious, cognitive science follows Darwin's anti-idealistic methodological lesson and proceeds *bottom-up*, attempting to reconstruct how the complex psychological functions underlying the adult self-conscious mind evolve from the more basic ones. This attempt does not appeal to our introspective self-

import the personal level at the sub-personal, in order to get all the sub-personal bits to do what they are supposed to do» (Gardner, 2000, pp. 100–101).

[9] In this context, the phenomenological aspects are considered to play a role in the mental life only insofar as they can be explicated in representational terms. See Lycan (2008).

knowledge, but instead appeals to those disciplines that investigate the gradual construction of self-consciousness as introspective reflexivity (Jervis, 2007, p. 152).

In this bottom-up perspective, it becomes possible to distinguish different forms of consciousness, which range from the simplest environmental monitoring to sophisticated forms of self-monitoring.

First, studies in cognitive ethology and developmental psychology tell us that neither infants under one year of age, nor most animals have the slightest idea – not even a confused one – of their own existence. They are conscious in the sense that they are able to form a series of representations of objects and operational plans of action, and hence to interact with persons and things in flexible but not self-conscious ways.

Second, some species take a step beyond the basic interactive monitoring of the environment that characterizes the simple consciousness of all animals. Great apes like chimpanzees, and in our species infants from 15–18 months of age, can be said to attain a state in which they are able to make a clear distinction between their own physical bodies and the surrounding environment. (More precisely, they first become capable of physical self-monitoring, i.e., focusing attention on the material agent as the (physical) executor of actions; and then their bodily self-monitoring comes to completion as the objectivation of a *proper body*, and thus as a rudimentary self-consciousness.)

Finally, it is only in human species, and only after the age of 3 or 4, that some unconscious psychological functions come to self-present themselves in accordance with the modes of self-conscious subjectivity. This is human consciousness in the traditional sense: self-consciousness as introspective recognition of the presence of the virtual space of the mind, separated from the other two primary existential spaces, i.e., the corporeal and extracorporeal spaces (see Jervis, 2007, p. 153).

By unearthing the non-primary but derived, constructed and partial character of self-consciousness, the cognitivist bottom-up approach can be regarded an *anti-phenomenology*, i.e., a critique of the subject, of its alleged givenness. The term "anti-phenomenology" was coined by Paul Ricoeur, who used it to define Freud's methodological approach. Ricoeur calls this approach «an *epoché* in reverse» (1970, p. 118). Freud's inquiry into the unconscious is an *epoché* in reverse because «what is initially best known, the conscious, is suspended and becomes the least known» (*ibid.*). 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. 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 (Jervis, 2011, pp. xxxi-xxxii). In contrast, cognitive science, fortified by a consciousness-independent concept of intentionality, rightly qualifies as an anti-phenomenology.

This allows us to estimate all the distance that separates the new cognitivist mentalism from the "consciousness-centric" mentalism that characterized the early experimental psychology, and from which the Freudian theory of the unconscious failed to disentangle itself. Under the influence of positivism, the introspectionist psychologists reified subjectivity. In most cases the 19th century experimental psychology did not understand consciousness in an experiential or subjective sense, but as an objective field, within which it was supposed to be possible to break down mental contents, viewed as measurable objects. As an antidote against the positivistic attempt to reify phenomenological experience, information-processing psychology provides us with a repertoire of tools to penetrate the nature of self-conscious subjectivity, making it possible to conceive phenomenological data not as tangible and measurable objects, but as the result of the self-presentation of unconscious psychobiological functions.[10]

4. Disunity and Opacity

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.[11] This theme

[10] The term "psychobiological function" points to my endorsement of teleofunctionalism, according to which «what makes a given type of mental state the type that it is, is its distinctive job or function within its subject's psychobiology» (Lycan & Neander, 2008).

[11] Although Freud does not offer an account of self-deception as such, his writings reveal very important characteristics of it that are not acknowledged by his "analytic" interpreters. See, e.g., Hållén (2011), who discusses self-deception in the context of Freud's writings and criticizes Davidson's and Gardner's analyses of the phenomenon.

can be considered the "strength" of Freud's conception of the unconscious.[12] A legacy, however, that can be capitalized provided one is willing to replace Freud's personal-level notion of dynamic unconscious with the new unconscious of neurocognitive sciences.

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*), Freud writes, the ego (*das Ich*) «appears to us as something autonomous and unitary, marked off distinctly from everything else» (1930/1962, 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 the 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» (Jervis, 2011, p. 43).

Today many behavioral, neuroimaging and computational investigations offer robust evidence for the composite, non-monadical nature of the mind-brain. In particular, since the early 1980s a *modularist* conception of the mind-brain has loomed large in psychology and neuroscience. The concept of modularity is to be placed in the framework of the crisis of the "pyramidal" conception of the mind, historically associated with the hierarchical conception of the cerebral functions dating back to the 19th Century. Against this view of mental life as a homogeneous and hierarchically-ordered field, ruled by consciousness and rationality, Noam Chomsky and David Marr have envisioned – in the wake of R. Mountcastle, D. Hubel and T. Wiesel's studies on the specializations of neurons – a less unitary, homogeneous, and hierarchically-ordered mind: its structure is *modular*, consisting of a bunch of distinct subsystems, that perform highly specific functions independently of each other (see Carruthers 2006).

Thus the neurocomputational architecture of our minds is composite and de-centralized, not monadic; and its appearing to consciousness as unitary is –

[12] This point is made by Jervis (2007, pp. 149–50). On Jervis' reconstruction of Freud's theory of the unconscious, see Marraffa (2011a,b).

as Freud suggested – a primary self-deception. To take just one famous example, in Dennett's narrative theory of personal identity the unitary consciousness of "self" is a short-lived "virtual captain" that occurs when a coalition of semi-independent, often domain-specific information processing mechanisms implemented in far-flung regions of the brain, has temporarily prevailed over other coalitions in the contest for the control of such activities as self-monitoring and self-reporting. Each of these short-lived phenomena is the 'me' of the moment, and they are connected to earlier fugacious selves by the autobiographical memory.[13] But then, "[i]f the temporary coalition of conscious states that is winning at the moment is what I am, is the self, each temporal chunk of 'self' is likely to be found in different parts of the brain from other such chunks and there will be many [neural correlates of consciousness] of unified consciousness in many different places" (Brook & Raymont, 2009, §7).

Freud's hypothesis that the presentation of the unconscious to introspective consciousness gives rise to deceptive beliefs about ourselves has found a rich source of evidence in the experimental social psychology literature on cognitive dissonance and self-attribution. Famously, in the experiments reviewed by Nisbett and Wilson (1977) the causes of the participants' behavior and attitudes (judgements, preferences and choices) were inaccessible motivating factors (e.g., subliminal cognitive inputs). However, when explicitly asked about the motivations (causes) for their behavior or attitudes, the subjects did not hesitate to sincerely affirm their plausible motives. The two psychologists explained this pattern of results by arguing that the subjects did not provide reports of real mental states and processes due to a direct introspective awareness; rather, they drew on repertoires of *rationalizations* seen as acceptable by mutual consent, and from time to time applied them, more or less stereotypically, to what needed to be justified.

Nisbett and Wilson's article was published in 1977. In the following thirty years the experimental literature on self-knowledge has increased substantially. Research in social and group psychology (e.g., Wilson, 2002; Wegner, 2002), in cognitive neuroscience (e.g., Hirstein, 2006) and cognitive neuropsychiatry

[13] Here I am following Brook & Raymont's (2009, §7) account of Dennett's view of the neural architecture of unified consciousness. The authors make clear that not any kind of autobiographical memory will be appropriate here; it must be «memory of the having, feeling, or doing of earlier experiences, emotions, actions, and so on» (Brook & Raymont, 2009, §5.2).

(e.g., Carruthers 2011) makes a very strong case for some version of a «symmetrical or self/other parity account of self-knowledge» (see Schwitzgebel, 2010, §2.1). According to the theory-theory version of this account, the attribution of psychological states to oneself (first-person mindreading) is an interpretative activity that depends on mechanisms that exploit the same folk theory of mind used to attribute mental states to other people. Such mechanisms are triggered by information about mind-independent states of affairs, essentially the target's behavior and/or the situation in which it occurs. The claim is, then, that there is a functional symmetry between first-person and third-person mentalistic attribution.

On this perspective, self-knowledge is not introspection insofar as this is construed as a direct access to the *causes* of our attitudes and behavior. In most cases of everyday life the explanation of the motives (being able to say "why") plays a *justificatory* role rather than a *descriptive* one. "Introspection" is then a misnomer for the capacity to explain one's behavior and attitudes *ex post* as the products of a rational and autonomous agent.

Moreover, Carruthers (2011) has extended this reappraisal of introspection beyond the causes of attitudes, to the attitudes themselves. According to Carruthers, we do not access propositional attitude events like judging and deciding via introspection; our only form of access to them is via self-interpretation, turning our mindreading faculty upon ourselves and engaging in unconscious interpretation of our own behavior, circumstances and sensory events like visual imagery and inner speech. Carruthers, therefore, develops a version of the symmetrical account of self-knowledge in which the theory-driven mechanisms underlying first- and third-person mindreading can count not only on observations and recollections of one's own behavior and the circumstances in which it occurs/occurred, but also on the recognition of a multitude of perceptual and quasi-perceptual events. Thus introspective consciousness comes out still more drastically downsized. True, agents have a sort of "perceptual" introspection. But this information is nothing but the raw material for an interpretative activity in which the access to the inner life is the access to an imaginary dimension generated by the folk-psychological theories driving the mindreading system.

Finally, Carruthers (2008) has put forward the hypothesis that Descartes' belief in the self-transparency of the mind reflects an innate feature of the human mind. According to this hypothesis, the mindreading system operates with a model of its own access to the rest of the mind that is essentially

Cartesian, assuming that subjects know, immediately and without self-interpretation, what they are experiencing, judging and intending. This assumption may have great heuristic value, greatly simplifying the mindreading system's computations. Moreover, as Wilson (2002) suggests, it may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image (a suggestion that anticipates the topic of the next section).

5. A Baconian Approach to Defense Mechanisms

Self-consciousness as introspective reflexivity is largely a theory-driven activity of re-appropriating the outputs of unconscious cognitive processing – this is the main point of the preceding section. Now what I want to emphasize is that such an activity is characterized by self-apologetic defensiveness: the description-narration of one's own inner life gets organized on the basis of the fundamental need «to construct and defend a self-image endowed with at least a minimal solidity» (Jervis, 1997, p. 33).

So we finally come to grips with the theme of defense mechanisms. But in view of neurocognitive sciences, the way in which Freud and his successors in the psychodynamic tradition have dealt with the study of psychological defenses must undergo a radical revision.[14]

We have already said that Freud's conception of the unconscious suffers due to an insufficient emancipation from the Cartesian model of the mind and the relationship between reason and passions. Descartes traced the errors of judgment and conduct back to the emotional, visceral, impulsive-instinctual, "animal" sphere of the body – this allowed him to safeguard the assumption of a primary (and for him transcendent) principle of human rational awareness. This ideology persists in non-specialist culture in the present day. The Cartesian faith in reason as producer of truth, the idea that what is clear and

[14] The notion of psychological defense is a psychoanalytic notion par excellence, whereas self-deception is a classical philosophical topic. Nevertheless, as McKay, Langdon and Coltheart (2009) rightly point out, defense mechanisms typically involve self-deception. Rationalization is a good example. The classic fable of the fox and the grapes, which nicely illustrates the "rationalization of disengagement", is a defensive maneuver through self-deception (see Elster, 1983). A variation of the sour grapes paradigm consists in rationalizing certain situations of intrapsychic conflict such as the cognitive dissonance investigated by Festinger in the 1960s, which illustrates the rationalization of "engagement".

distinct cannot be false, and that errors are essentially a sort of derailment due to drive-visceral interferences, is implicit also in Freud's system of thought.

But the Cartesian conception of error pays heavy tribute to philosophical predecessors of the modern era. It had already found an implicit refutation in Francis Bacon's work, which traces the errors of judgment and conduct back to the forms of doing and knowing that are peculiar to the psychological essence of human beings. In Bacon, contrary to Descartes, the conscious and rational mind naturally produces errors: the human understanding, he writes, «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» (1620/2000, p. 41). We could say, in current terms, that Bacon sees the mind's errors, illusions, and self-deceptions as intrinsic to the ordinary cognitive-affective processes. It is on these grounds that he claims the necessity of a system of tests through which our spontaneous tendency to make errors is "dug out" and rectified by the method of research, on the base of a rigorously empiristic methodological principle.[15]

It is this Baconian perspective that has been taken by research traditions such as psychology of thought and social psychology. Thus, for example, social psychology tells us that stereotypes, the dynamics of prejudice, the structurally unreliable or diverting nature of many programmatic and principled avowals, are structures of bad faith which originate from cognitive mechanisms underlying the etiology of social attitudes. In such a perspective, then, 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, consisting of passions, instincts, emotions, and which can be clearly demarcable from the workings of our self-conscious rationality.[16] Now self-deception is a natural inclination of the human mind, a property inherent to belief-formation mechanisms (see, e.g., Bayne and Fernández, 2009, pp. 5–6).

This gives rise to a *reinforcing overturning* of the psychodynamic questioning about defenses. Now «the aspects of ambiguity, self-deception, and [...] sufferance of human life» can no longer be conceived as «interferences

[15] See Jervis (1993, pp. 122–123), who refers to Paolo Rossi's works on this topic (see, e.g., Rossi 1968).

[16] On the other hand, the folk concept of emotion is not a natural kind, i.e., a category that groups together a collection of objects whose properties are correlated by virtue of a causal mechanism that makes it possible projection and induction. On this point, see Griffiths (1997).

that are restrictively connected to affective and emotional factors (and hence negatively affecting a self-conscious rationality safeguarded as primary)»; they are to be seen as aspects «globally constitutive of the mind and behavior» (Jervis, 1993, p. 302). What needs to be explained, then, is not «how and why some defense mechanisms exist, but rather how all the structures of knowledge and action are by themselves, integrally, a matter of defenses» (Jervis, 1993, p. 301). In short, defense mechanisms are mechanisms that permit us to think and act. Although their most manifest function is that of protecting from anxiety, defense mechanisms are the primary instruments for setting up order in the mind. Consequently, we are now able to capture something that is already in Freud but which the Cartesian model prevented him from thoroughly articulating: the defensive processes are something more than bulwarks against anxieties and insecurities that perturb the order of our inner life; actually, defense mechanisms are the very structure of the mind – the Freudian ego itself is a defense. Here are the roots of the clinical theme of the fragility of the ego, namely that intimate personal insecurity that seems to originate from insufficiencies in the primary relationship between mother and child (what Michael Balint termed "basic fault"). But the theme is much wider, and it has to do with a philosophical anthropology that is congruent with the ontology of neurocognitive sciences.

Then let us ask ourselves: who is the subject of a dynamic psychology based on the cognitive-science ontology of unconscious psychobiological functions? After undergoing the above-mentioned "reinforcing overturning", the ideas of the unconscious and defense mechanisms have no longer the function of downsizing the traditional image of a subject with a primary identity and force; on the contrary, they certify the non-existence of a human subject of that kind. What, more than anything else, defines the real human subject is its original lack of ontological consistency. Unlike Descartes' soul-like ontologically guaranteed consciousness-substance, the image of the subject that cognitive sciences deliver us is that of a multiplicity of functions that in presenting themselves to consciousness exhibit a "façade" made of representations of the self. But it is a façade that is inextricably marked by *bad faith*; that is, «it is something inauthentic and bi-dimensional, i.e., "shallow", which tends to pass itself off – in accordance with our insuppressible tendencies to self-deception – as the 'solid', or 'deep', structure of the person» (Jervis, 2011, p. 45).

These dynamics of the representations of the self are the dynamics of the *subjective identity*, namely the consciousness of the self as description of the

self. I know that I exist insofar as I know that I exist "in a certain way", as describable identity, constant through changes. This theme is well captured by William James: every day, at each awakening, I find again my own body and my own mind, namely myself as known identity – «Each of us when he awakens says, Here's the same old self again, just he says, Here's the same old bed, the same old room, the same old world» (James, 1890/1950, p. 334).

However, self-consciousness as finding oneself again as known identity, as feeling of being-here as being-here in a certain way, is a *precarious* acquisition, continuously constructed by the subject and constantly exposed to the risk of not being here. If the subject's self-description becomes uncertain, she soon feels that the feeling of existing vanishes. This can occur for various reasons: because of a sudden breakdown of self-esteem; on the occasion of unexpected emotional upheavals; in some cases of psychoses or loss of memory; when the continuity of the tissue of our sociality is broken, as it can happen when one is suddenly thrown in some dehumanizing total institution (Jervis, 2011, pp. 131–132).

It is therefore the precariousness of this description of identity that makes intelligible the primary defensiveness of the self-constructing subject. The human subject constitutes itself as a repertoire of defensive maneuvers that must cope with its ontological insubstantiality. It could be said that the mind achieves its appearance of unity in the act of mobilizing tricks against the threat of its breaking down. And it is worth noting that such an activity – aimed to defend one's own self-describability and, indissolubly, the cohesiveness of one's own self-conscious consistency – is not restricted to an individual, psychodynamic dimension, i.e., to the intrapsychic defenses and the interpersonal maneuvers to which we appeal in the relationship with other people and our social environment. For it also has a collective, anthropological dimension, where the defenses consist in the construction of a system of references, in part symbolic and ritual, which give meaning to one's own being in the world.[17]

[17] See Jervis (2011, p. 92), who is building on Ernesto de Martino's seminal work on the "crisis of presence". This is a breakdown in the sense of self that occurs in the confrontation with death, in cases of psychological dissociation, alienation, and «loss of subjectivity, i.e., of one's ability to act on the world rather than simply to be a passive object of action» (Saunders, 1993, p. 882). According to de Martino, overcoming the crisis of presence is the fundamental task of culture.

6. Conclusion

Self-deception can be seen as a paradox of rationality only within the framework of the Cartesian conception of a self-transparent, unified and integrated self. Once we abandon the Cartesian theory of the subject, and invoke the subpersonal framework of neurocognitive sciences, self-deception is, in its primary form, a way of alluding to a mismatch: the description of the self as a description of identity is irreducibly out of phase, i.e., heterogeneous, with respect to the much more composite reality of the neurocognitive unconscious. Our mind is not self-transparent, i.e., essentially it eludes us, and also "deceives" us; and it deceives us just starting from its pseudo-transparency and consciential pseudo-unity. The mind contains non-truth-tropic cognitive mechanisms that generate the reassuring effect of a unitary egoic subjectivity that is master of the contents of consciousness. This effect is a "façade" whose deceptive character will be denied if human beings must feel their own autonomy, and thus experience themselves as persons. Or equivalently, the activity of narrative re-appropriation of the products of the unconscious cognitive processing is ruled by the fundamental need «to construct and protect a self-image endowed with at least a minimal solidity, and that is, in practice, solid enough to confirm to ourselves that we exist without dissolving ourselves» (Jervis, 1997, p. 33). This is the framework within which we can understand the construct of defense mechanisms, and with it all variety of self-deception.

REFERENCES

Bacon, F. (1620/2000). *The New Organon*. Edited by L. Jardine, & M. Silverthorne. Cambridge: Cambridge University Press.

Bayne, T., & Fernandez, J. (2009). Delusion and Self-Deception: Mapping the Terrain. In T. Bayne, & J. Fernandez (Eds.), *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. New York: Psychology Press, 1–20.

Brook, A., & Raymont, P. (2010). The Unity of Consciousness. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall2010/entries/consciousness-unity/>.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Carruthers, P. (2011). *The Opacity of the Mind*. Oxford: Oxford University Press.

Cottingham, J. (1988). *The Rationalists*. Oxford: Oxford University Press.

Davidson, D. (1982). Paradoxes of irrationality. In R. Wollheim, & J. Hopkins (Eds.), *Philosophical Essays on Freud*. Cambridge: Cambridge University Press, 289–305.

Davidson, D. (1998). Who is fooled? In J. Dupuy (Ed.), *Perspectives on Self-Deception*. Cambridge: Cambridge University Press, 1–18.

Descartes, R. (1641/1988). Author's replies to the second set of objections. In J. Cottingham, D. Murdoch, & R. Stootho (Eds.), *The Philosophical Writings Of Descartes*. Cambridge: Cambridge University Press, vol. II, 93–120.

Dennett, D. C. (1991). *Consciousness Explained*. New York: Little, Brown & Co.

Dennett, D.C. (1993). Review of J. Searle, The Rediscovery of the Mind. *The Journal of Philosophy*, *60*, 193–205.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Elster, J. (1984). Managing to deceive ourselves. *The Times Literary Supplement*, *4261* (November 30), 1388.

Freud, S. (1930/1962). *Civilization and its Discontents*. Translated and edited by J. Strachey. New York: Norton.

Gardner, S. (1999). Psychoanalysis, contemporary views. In R.A. Wilson, & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge (MA): MIT Press, 683–685.

Gardner, S. (2000). Psychoanalysis and the personal/sub-personal distinction. *Philosophical Explorations*, *3*(1), 96–119.

Griffiths P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: Chicago University Press.

Hållén, E. (2011). *A Different Kind of Ignorance: Self-Deception as Flight from Self-Knowledge*. Uppsala University, PhD dissertation.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge (MA): MIT Press.

James, W. (1890). *The Principles of Psychology*. New York: Dover, 1950.

Jervis, G. (1993). *Fondamenti di psicologia dinamica*. Milan: Feltrinelli.

Jervis, G. (1997). *La conquista dell'identità*. Milan: Feltrinelli.

Jervis, G. (2007). The unconscious. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the Mind*. Berlin: Springer, 147–158.

Jervis, G. (2011). *Il mito dell'interiorità*. Tra psicologia e filosofia. Edited by G. Corbellini, & M. Marraffa. Turin: Bollati Boringhieri.

Johnston, M. (1988). Self-deception and the nature of mind. In B.B. Mclaughlin, & A. Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 63–91.

Kandel, E. (2005). *Psychiatry, Psychoanalysis, and the New Biology of Mind*. Arlington (VA): American Psychiatric Publishing.

Laplanche, J., & Pontalis, J.-B. (1967). *Vocabulaire de la psychanalyse*. Paris: Presses Universitaires de France.

Livingstone Smith, D. (1999). *Freud's Philosophy of the Unconscious*. Dordrecht: Kluwer.

Lycan, W. (2008). Representational Theories of Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall2008/entries/consciousness-representational/>.

Lycan, W.G., & Neander, K. (2008). Teleofunctionalism. *Scholarpedia*, *3*(7), 5358.

Manson, N. (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*. London: Routledge, 148–168.

Manson, N. (2003). Freud's own blend: functional analysis, idiographic explanation, and the extension of ordinary psychology. *Proceedings of the Aristotelian Society*, *2*, 179–195.

Marraffa, M. (2011a). Precariousness and bad faith. Giovanni Jervis on the illusions of self-conscious subjectivity. *Iris*, *3*(2), 171–187.

Marraffa, M. (2011b). Jervis e la genealogia nascosta della coscienza umana. In G. Jervis, *Il mito dell'interiorità. Tra psicologia e filosofia*. Edited by G. Corbellini, & M. Marraffa. Turin: Bollati Boringhieri, XI–LVI.

Marraffa, M. (2011c). Jervis, De Martino e il mito dell'interiorità. *Rivista di Filosofia*, *2*, 241–259.

Marraffa, M. (forthcoming). Troubles with self-consciousness. Jervis on introspection and defense mechanisms. *Medicina nei secoli*, *23*(1), 2012.

McKay, R., Langdon, R., & Coltheart, M. (2009). "Sleights of mind": Delusions and self-deception. In T. Bayne, & J. Fernandez (Eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Hove: Psychology Press, 165–185.

McWilliams, N. (1994). *Psychoanalytic Diagnosis*. New York: Guilford Press.

Mele, A.R. (1997). Real Self-Deception. *Behavioral and Brain Sciences*, *20*, 91–102.

Mele, A.R. (2009). Delusional Confabulations and Self-Deception. In W. Hirstein (Ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy*. Oxford: Oxford University Press, 139–157.

Nisbett, R.E., & Wilson, T.D. (1997). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

O'Brien, G., & Jureidini, J. (2002). Dispensing with the dynamic unconscious. *Philosophy, Psychiatry and Psychology*, *9*(2), 141–153.

Pears, D. (1982). Motivated irrationality, Freudian theory and cognitive dissonance. In R. Wollheim, & J. Hopkins (Eds.), *Philosophical Essays on Freud*. Cambridge: Cambridge University Press, 279–288.

Pears, D. (1984). *Motivated Irrationality*. Oxford: Oxford University Press.

Ricoeur, P. (1970). *Freud and Philosophy. An Essay on Interpretation*. New Haven and London: Yale University Press.

Rossi, P. (1968). *Francis Bacon: From Magic to Science*. London: Routledge.

Sage, J. (forthcoming). The Evolutionary Basis of Self-Deception. <http://www4.uwsp.edu/philosophy/jSage/Sage%20Evolutionary%20Basis%20of%20Self-Deception.pdf>.

Sartre, J.-P. (1943). *L'être et le néant*. Paris: Gallimard.

Saunders, G. R. (1993). "Critical Ethnocentrism" and the ethnology of Ernesto De Martino. *American Anthropologist*, *95*(4), 875–893.

Schwitzgebel, E. (2011). Introspection. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2010/entries/introspection/

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge (MA): MIT Press.

Wilson, T.D. (2002). *Strangers to Ourselves*. Cambridge (MA): Harvard University Press.