RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2014.0016

ISSN 2039-4667; E-ISSN 2239-2629
Vol. 5 (2014), n. 2, pp. 207-220

Studi

The Unconscious, Self-Consciousness, and Responsibility

Massimo Marraffa

Ricevuto il 18 giugno 2013; accettato il 4 maggio 2014

Abstract In this article I argue that (1) introspective self-consciousness is an activity of narrative re-appropriation of the products of the cognitive unconscious; and (2) this activity has an essentially self-defensive character, being ruled by the primary and universal need to construct and protect a subjective identity whose solidity is the ground of the intrapsychic and interpersonal balances of human organism. Finally, in this framework firmly based on psychological sciences, I reconsider John Locke's link between responsibility and self-consciousness.

KEYWORDS: Identity; Responsibility; Self-consciousness; Self-narration; Unconscious

Riassunto *Inconscio, autocoscienza e responsabilità* – In questo articolo sostengo che (1) l'autocoscienza introspettiva è un'attività di riappropriazione narrativa dei prodotti dell'inconscio cognitivo; e (2) questa attività ha una natura essenzialmente difensiva, essendo governata dal bisogno universale e primario di costruire e proteggere un'identità soggettiva la cui solidità è il fondamento degli equilibri intrapsichici e interpersonali dell'organismo umano. Infine, in questo quadro saldamente fondato sui dati delle scienze cognitive, riconsidero il nesso istituito da John Locke fra responsabilità e autocoscienza.

PAROLE CHIAVE: Identità; Responsabilità; Autocoscienza; Narrazione di sé; Inconscio

Locke on self-consciousness and responsibility

According to Locke, the concept of person does not refer to an essence but rather to a psychosocial attribute that is assigned to those subjects who possess a specific set of psychological capacities.[1] This is in agreement with the most common legal language, which suitably talks about "natural persons" and similarly about "legal persons", thus pointing out something precise, i.e., the presence of an agent or subject who, in virtue of one's intrinsic characteristics, is fully able to perform such acts as buying a 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. So she is imputable as well: if she committed a crime, she knew very well what she was doing. The concept of person there-

M. Marraffa - Dipartimento di Filosofia, Comunicazione e Spettacolo, Università di Roma Tre -
via Ostiense, 134 - 00154 Roma (✉)
Email: massimo.marraffa@uniroma3.it

fore rests on that of personal responsibility; and 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, not even an old man who suffers from arteriosclerosis and 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 owns a collection of psychological capacities. It is someone who is able to form imaginary test scenarios 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, i.e., an inner virtual space in which thoughts and affects as "private" events can be situated. Only someone who has enough access to one's interiority (to oneself as objectified in the introspective consciousness of the self) can appropriate «Actions and their Merits».[2]

In Locke, therefore, an individual is a person only insomuch as she can reflectively appropriate her actions and their meaning – an appropriation that originates from «that consciousness which is inseparable from thinking».[3] Locke also realizes that this very awareness is the ground of the sense of identity. What is really 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: «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.[4] But if the person is determined by consciousness, by what is consciousness determined?

Locke relies on 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; actually it is still a sort of secularized soul. Despite the philosopher's good intentions, it is described as a sort of essence. For all that, Locke's consciousness is still given a priori.

A different kind of consciousness can be found in psychological sciences: something that is *constructed* during life, that emerges from the multifarious qualities of the body and of human existence. And it is from this standpoint that now I will reconsider Locke's link between responsibility and self-consciousness.

The Freudian Unconscious

Freud's originality does not consist in the discovery of the unconscious: the terms "subconscious" and "unconscious" were already in currency in the last decades of the 19th century, introduced to explain phenomena (e.g., convulsive "great" hysteria, dissociative fugue or multiple personality disorder) that could hardly be reconciled with the Cartesian consciousness-dependent conception of mind that was shaping the early experimental psychology.[5] Rather Freud's originality consists in developing the concept of unconscious in two particular directions.

In the first place, Freud puts forward the idea of a *sexuality* of the unconscious. At the heart of the unconscious there are "energies", "forces" that Freud calls *Triebe* (instinctual drives) – first and foremost, the sexual instinctual drive, or *libido*. 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 the 19th century middle class ethics – a model that rested on the assumption of a full responsibility of the individuals towards an inner life made of conscious and self-transparent intentions.

Such a revision was fostered, on the one hand, by the Darwinian naturalism and the medical biologism of the 19th century; on the other, by an anthropology of "the crisis of Reason" which, originated from Romanticism and the sceptical thought of the past centuries (above all Hume's), had found its main theorists in Schopenhauer and Nietzsche. However, Freud 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 a bad administration of the relationships with the unconscious, one that resulted in wrong forms of self-repression. In this perspective, the psychoanalytic therapy offered the attractive perspective of a better managing 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.[6]

In a scientific perspective, however, the conceptualization of sexuality in terms of instinctual drives is definitely the most timeworn part of Freud's work. (And it is not a minor shortcoming since the bioenergetic model of the mind is the main doctrinaire premise of Freud's psychoanalysis.) The debate on the concept of instinct with its variations (tropisms, reflexes, drives etc.) goes with all the history of psychology. Attacked already 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 both on the biological front, by virtue of the study of behavior in terms of signals due to the British school of ethology,[7] and on the experimental front, at first in relation to the development of the studies on the mechanisms of learning, and then for 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, the psychological functions (a concept that Freud did not possess) are defined in terms of signals and information. But also in the psychoanalytic field, already in Michael Balint, in the 1930s and 1940s, we find an implicit crisis of the centrality of instinctual drive in the theory of object relations.

According to this trend in psychoanalysis, the quest for the "object" is not secondary to the need to discharge the libidinal drive, it is primary; this drastically downsizes the role of instinct as energy. The criticism of the concept of instinctual drive becomes explicit with John Bowlby's theory of attachment. Finally, the most systematic and radical attack against Freud's idea of instinct is launched in the United States, 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.[8]

The second main feature of the Freudian concept of the unconscious is that "unbearable" mental contents are unconscious in that *repressed*, viz. actively excluded from awareness owing to the unconscious activating of defensive mechanisms. This concept, too, is timeworn. For it has now become clear that the phenomenon that Freud called "*Verdrängung*" – totally removing or cancelling the memory of a traumatic event (prototypically, an episode of sexual molestation in infancy) from our conscious minds (and irreversibly, unless making appeal to specific techniques like hypnosis or psychoanalytic theory) –, if it even 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.[9]

After Freud, however, a weaker sense of "repression" established itself, even at a commonsense level. This is a meaning that we find in quite usual sentences like, for example, "I remembered the date only when it was already too late". If someone said, as now it is very common to say, that I "repressed" the recollection of that date, what he would mean to say is not that I cancelled the date from memory but rather that I have temporarily put it aside (in an *interested* manner: I did not want to remember).

This weaker sense of repression is highly relevant to the extent that it is consistent with a view of consciousness that is different from Freud's. On his view, on the one side there is consciousness (well separated from the unconscious), on the other, the "stumbles" of consciousness (caused by the unconscious's infiltrating into consciousness). Such stumbles occurred only in a few exceptional or anomalous cases like, precisely, repressions (in the strong sense).[10] 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, viz. by a multitude of defensive strategies that are very akin to repressions (in the weak sense).[11]

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 still living matter, that of *bad faith*: our mental processes are permeated by «a self-apologetic defensiveness or rather, a systematic tendency towards self-deception within our everyday thought processes».[12] This critical theme – the tendency of the mind to forge self-serving illusions – is the "strength" of Freud's concept of the unconscious.

Freud's hypothesis is then that our consciousness somehow deceives us by providing us with illusory immediate beliefs about ourselves. In other words, the consciousness of self flatters us with the quite apparent presentation of qualities that are additional to the reality of the way in which the mind works. First and foremost, there is a mismatch between the composite, non-monadical character of the mind and its unitary phenomenology. In the *Ichgefühl*, Freud writes, the ego «appears to us as something autonomous and unitary, marked off distinctly from everything else».[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 in continuity without any sharp delimitation and for which it serves as a kind of «façade».[14]Accordingly, the ego is both the organization of the psyche (i.e., the partial structuration of the disparate functions of the mind) and the apparatus that has, among its various tasks, that of setting up a complex self-deception, viz. the narration of the self as an imaginary entity which is primary, unitary, free, rational, master of the person.

Freud's systematic doubt about the traditional claims of self-legitimation of consciousness opens a crack in self-consciousness; a crack that – as we will now see – becomes a ruinous landslide in psychological sciences.

The Cognitive Unconscious

As I just said, the critical theme of bad faith is the strength of Freud's concept of the unconscious. By contrast, the relationship between consciousness and the unconscious, as this unfolds in Freud's theory of repression, is the clue of the main difference between the psychoanalytic unconscious and the unconscious and subpersonal 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 about the mind – actually it could be said that the latter is a largely psychoanalytic culture.[15] And 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, which was challenged by Freud.

If the Victorian 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 folk psychoanalytic 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.[16]

Thus the folk psychoanalytic culture of the mind 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 this culture holds what is the most evident limitation of Freud's view of the unconscious: 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 view.[17]

As a result, the Freudian unconscious comes to be just an enlargement, or extension, of a psychology – folk psychology – hinged on the idea of a person who is able to have conscious mental experiences. As Laplanche and Pontalis noticed, in Freud's second topography the model is no longer one borrowed from the physical sciences as it was in the case of first topographical conceptualisation of the psychical apparatus:

> 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 towards 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.[18]

In brief, psychoanalysis turns out to be a *personal* psychology that is masked as *subpersonal* psychology.

A response to this difficulty of psychoanalysis will consist in opposing to the Freudian unconscious the "new unconscious" that has emerged from a variety of disciplines that are broadly part of cognitive science.[19] This is a level of analysis that aspires to be genuinely subpersonal: the information-processing level, wedged between the personal sphere of phenomenology and the subpersonal domain of neurobiological events. Such level no longer takes consciousness as an unquestionable assumption, as a non-negotiable given fact; the concept of cognitive unconscious is no longer patterned, as in Freud, after the concept of conscious mind.

Rather cognitive science's subpersonal processes show different features 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.

This claim is to be calibrated bearing in mind that in some cases also the cognitive-science unconscious processes are a little too akin to the idea that is intuitive to the folk – some cognitive-science models, and specifically Jerry Fodor's computational-representational theory of mind, tend to reproduce the operation of the conscious thought processes. So, 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 still that cognitive processes (including the perceptual ones) can be assimilated to deductive chains. On the whole, however, it can be affirmed that, because of the very way of conceiving the mind in cognitive science as something halfway between the person and the brain, the cognitive unconscious does not faithfully reflect the conscious level; and that the models of the unconscious more adhering to the structure of awareness are likely to belong more to the past of cognitive sciences than to their present.[20]

A bottom-up approach to self-consciousness

Cognitive sciences, therefore, challenge the traditional nexus between consciousness and intentionality, thus opening a conceptual space in which to build a theory of the "non-derived" unconscious, viz. a theory that no longer arrives at the unconscious by subtraction from consciousness. In Dennett's terms, first one develops a theory of intentionality that is independent of and more fundamental than consciousness, a theory that makes no distinction between the various forms of unconscious representational mentality.[21] Then, one proceeds 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 would have it, the foundation of everything mental. In short, first intentionality, then consciousness.[22]

Viewing consciousness no longer as something that explains, but rather as something that needs to be explained, analyzed, dismantled, is also in full agreement with Darwinian naturalism. In asking how consciousness, rather than the unconscious, is possible, the cognitive scientist fully endorses Darwin's methodological approach, which, assuming the continuity between animal and human minds, pursues the study of consciousness by virtue of a bottom-up strategy, i.e., reconstructing how the complex psychological functions underlying the adult self-conscious mind evolve from more basic ones.

The bottom-up approach allows us to draw a clear-cut distinction between object-consciousness and self-consciousness. For data from cognitive ethology and developmental psychology provide grounds to hold that infants under one year of age 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 ways, but this occurs automatically, prereflexively (nonconceptually), without any cognition of a bodily or "inner", experiential space.[23]

Few species take a step beyond this basic interactive monitoring of the environment. Great apes like chimpanzees, and in our species infants from 15-18 months of age, can be said to attain awareness of their bodies as unitary sources of their actions and their gazes – as measured by the mirror self-recognition test. But note: 15-18 months infants have come to grips with the subjective-objective space of the body but not yet with the virtual space of the mind. They are not able to objectify their own subjectivity knowing that it is *their own* subjectivity, in the same way as at "one and a half year of age" they had objectified their own body knowing that was *their own* body. For instance, three-year-olds are not yet able to understand dreams as non-real, private, psychological occurrences; instead, they consider them as either real events or visions "sent from outside", which crowded their bedrooms.[24]

It can be supposed that at an early stage human bodily self-consciousness, such as that of the chimpanzee, is structured by a non-conceptual representation of the physical self, but very soon it begins to be mediated by the verbal exchange with the caregiver. In other words, in our species the chimpanzee-style, purely bodily self-awareness is almost immediately outstripped and encompassed by a form of descriptive self-consciousness that is strictly linked to linguistic tools and social cognition mechanisms.

Consequently, around the age of 3 or 4 years something occurs that can be observed only in the human species: the child discovers that she has an "inner life", i.e., she becomes able to identify and objectify her own subjectivity. Here the lived subjective experience takes as its object not only the outer world (as happens in all animals), not only the bodily world (as happens in chimpanzees and 15-18 month old children), but also *itself*. This is self-consciousness as *introspective* recognition of the presence of the virtual inner space of the mind, separated from the other two primary experiential spaces, viz. the corporeal and extracorporeal spaces; the consciousness as identity of person which, on Locke's view, grounds the notion of responsibility.[25]

The Interpretive Sensory-Access theory of self-knowledge

When this introspective consciousness is put under the magnifying lens of cognitive sciences, however, the question arises whether it is really "awareness" or rather a self-constructive story-telling that, as Freud saw it, is infused with self-deceptions and bad faith.

Freud's idea of a pervasive presence of self-deception in our inner life has found a rich source of evidence in the extensive cognitive dissonance and causal attribution literatures that have built up in experimental social psychology over the last fifty years. In Nisbett and Wilson's well-known 1977 review of these literatures, the participants' behavior was caused by motivational factors inaccessible to consciousness. However, when explicitly asked about the *motivations* (causes) of their actions, the subjects did not hesitate to sincerely affirm their plausible *motives*. Nisbett and Wilson 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 engaged in a "confabulatory" activity, i.e. they used a *priori causal theories* in order to develop reasonable but imaginary explanations of the motivational factors of their behaviors, judgments or decisions.[26]

Nisbett and Wilson's account of causal self-attribution in terms of theory-laden confabulatory activity is an exemplar of what Schwitzgebel termed "self/other parity account of self-knowledge", since the attribution of psychological states to oneself is seen as an interpretative activity that depends on mechanisms that exploit theories that apply to the same extent to ourselves and others.[27] Such theory-driven mechanisms take as input information about mind-external states of affairs, essentially the target's behavior and/or the situation in which it occurs.

On this perspective, then, introspection, insomuch as it is construed as a source of knowledge of the (multifactorial) aetiology of our judgments, decisions and behavior, is an illusion. In its stead we find the theory-driven 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, giving reasons for what has been done ("being able to say why") plays a *justificatory* role rather than a *descriptive* one.

It is to be noticed, however, that the self/other parity account is never suggested as an *exhaustive* theory of self-knowledge; for some margin is always left for some sort of *direct* self-knowledge.[28] Nisbett and Wilson, e.g., draw a sharp distinction between *process* and *content*, i.e., between the causal processes underlying judgments, decisions, emotions, sensations and those judgments, decisions, emotions, sensations themselves. Subjects have direct access to this mental content, and this allows them to know it «with near certainty».[29] By contrast, they have no access to the cognitive processes that cause behavior. However, insofar as the two psychologists do not offer any hypothesis about this alleged direct self-knowledge, their theory is incomplete.

Peter Carruthers tried to bridge this gap by developing an "Interpretive Sensory-Access" (ISA) theory of self-knowledge.[30] This theory is well in line with the global workspace model of the human neurocognitive architecture (first postulated by Bernard Baars),[31] which posits a range of perceptual systems that broadcast their outputs to a suite of concept-using consumer systems. Among these there is a mindreading system which, driven by a folk-psychological theoretical framework, produces higher-order, metarepresentational, beliefs about the mental states of others and of oneself.

The mindreading system, then, has access to all sensory information broadcast by our perceptual systems; and hence it can have a non-interpretive ("recognitional") access to one's own sensory and affective states. But the system cannot directly self-attribute "thoughts" (i.e., propositional-attitude events). For the latter are not globally broadcast but are the output of conceptual consumer systems; and there aren't any causal pathways from the outputs of these systems to mindreading, which would be necessary to allow introspective access to one's thoughts.

As a result, the mindreading system must exploit the globally broadcast perceptual information, together with some forms of stored knowledge, to infer the agent's thoughts, precisely as it happens with the reading of other minds. Thus 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 (e.g., imagery or sentences in inner speech).

Carruthers offers a great number of arguments for his ISA theory: considerations concerning the evolutionary role of mindreading and the literature on metacognition (see below); evidence from the literature on confabulation that allows him to apply the self/other parity account of self-knowledge to thoughts; data from psychopathology that refute the above-mentioned hypothesis of a dissociation between self-attribution and other-attribution. Thus he develops a sophisticated version of the self/other parity account of self-knowledge in which the theory-driven mechanisms underlying mentalistic self-attribution and other-attribution can count not only on the observation/recollection of one's own behavior and/or the circumstances in which it occurs/occurred, but also on the recognition of a multitude of perceptual and quasi-perceptual events.

All this delivers us a drastically debunked conception of our inner life. Except for perceptive and quasi-perceptive events, there are no *conscious* mental phenomena; there is definitely no phenomenology of thought (of such events as judging, intending, or deciding). 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.

It is be noted that if this eliminative claim about conscious thought is well grounded, we have here a very strong constraint on the construction of a theory of moral responsibility congruent with the findings of cognitive sciences: the existence of conscious intentional mental states cannot be among the theory's commitments. Thus, to make only one example, let us consider the theories of the *real self*.[32] These theories claim that an agent can be held responsible exclusively for those actions that have been caused by psychological states reflecting its identity as practical agent. But if – as it seems to be necessary – the psychological states that define the agent's real self are the conscious ones, the elimination of conscious thought implies the non-existence of the real self.[33]

The armchair moral philosopher could look at this constraint with impatience, rebutting that whereas the literatures of biological and psychological sciences are constitutive of the descriptive ethics, their relevance for normative ethics and meta-ethics appears to be much more restricted; and this is because these two areas of study require the use of a normative conceptual apparatus whose reducibility to naturalistic categories is highly controversial. But to this one can reply that the right acknowledgement of the specificity of the normative dimension should not go to the point of concealing its intimate dialectic with the descriptive sphere. For any moral statement is not exclusively prescriptive but also contains factual beliefs, which can be true or false.

Therefore, the normative use of the cognitive-science findings can consist in a critical examination of the descriptive ingredient of our moral statements, revealing in some cases its close connection to conceptions of human nature that today we are able to unmask as the fruits of the imagination of philosophers and theologians.[34]

The defensive nature of self-consciousness

Under the lens of cognitive sciences, the Lockean self-consciousness turned out to be not a direct access to inner life but a theory-driven activity of narrative reappropriation of the products of the cognitive unconscious. Now our focus will be on the intrinsically defensive nature of this self-narration.

"Let us go back to" the ontogenesis of introspective self-consciousness. With the development of social cognition and linguistic-narrative competence, the physical, bodily self-description becomes psychological, introspective. And this mentalistic self-description is likely to take shape in the act of turning on oneself the capacity to mindread other people – i.e., the understanding of other minds both ontogenetically precedes and grounds the understanding of our minds.

As we have seen, Carruthers has made a strong case for the claim that third-person mindreading has a functional and evolutionary priority over first-person mindreading. The mindreading system can be said to be focused outwards on the world rather than inwards on the agent's own mental states.[35] And 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 two motivational macro-systems: the first committed to self-assertiveness and competition,[36] the second aimed to pro-sociality and cooperation.[37]

In this perspective, one virtue of Carruthers' model lies in its explanatory parsimony from an evolutionary point of view; for it posits a single phylogenetic route for both third person and first person mindreading. If, as the ISA model holds, first person mindreading results from turning one's third person mindreading capacities upon oneself, the emergence of the former will be a by-product of the evolution of the latter. By contrast, theories to the effect that first person and third person mindreading are subserved by two (or more) neurocognitive mechanisms bear an explanatory burden, because then there should plainly be a distinct evolutionary story to be told about the emergence of each.[38] And to date we do not have a plausible hypothesis about what kind of evolutionary pressure can account for the emergence of first person mindreading mechanism(s).[39]

The ISA model holds that third-person mindreading has a functional and evolutionary priority over first-person mindreading, but it does not predict that the former is developmentally prior to the latter.[40] However, there are good reasons for thinking that the inner experiential space is constructed *outward-in*; and that this occurs around the age of 3-4 years and in an interpersonal context, viz. in the relationship with the caregiver.

More precisely, it can be supposed that one of the factors that give rise to inner life is a component of the mindreading system that systematically reads behaviors of other people as actions driven by goals, purposes, intentions (and intentions with a positive or negative valence).[41] The question "What does *that* want to do?" (where "that" can refer to the mother or the home cat) is already asked in infancy and toddlerhood. And then, on the basis of this kind of questions, children begin to ask *also* what their own intentions are, what their own inner state is. This appropriation of themes that initially were only connected 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 a large part of simplest introspections are forms of learning emerging from the verbal stereotypes and rhetorics through which adults rename the intentions of others. A two year old, maybe because she is scared by her granny's cat, maybe as an act of defiance, gives the cat a boot; and here are the reconstructive judgements 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 had scared you, but perhaps that kitty was more scared than you". And so the child gradually learns – and 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.[42]

Note the connection between the construction of inner life and ethics, to which Locke had already drawn attention. Morality reinvents inner life from scratch: being bad and being good, having bad intentions and having good intentions, appear to the child the premise of imputability even before of responsibility. This permits to explain why, despite the above-mentioned verbal stereotypes and rhetorics actually contain a plea for responsibility, in our culture the sense of *responsible* appropriation of one's own actions – so I know that I could be objectively and legally responsible for a car accident even if I could not be able to identify in myself an intention to cause it[43] – is usually replaced by a more unclear and 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 intents is not taken into account by our intuitive psychology.

Introspective self-consciousness arises therefore in the child in a relationship with the caregiver that is made of words, descriptions, designations, evaluations of the person. Through the dialogue with the caregiver (and then with other social partners) the 3-4-year-old child builds itself by constructing its own identity, both objective (i.e., for others) and subjective (i.e., for itself). And following G.H. Mead's lesson, we can say that the identity-for-itself largely derives from the identity-for-others; namely, we see ourselves, and define ourselves, essentially introjecting the way in which others see and define us.

The child's inner experiential space gradually takes the form of subjective identity – i.e., the child gradually comes to experience himself as a person, to define himself as a certain kind of person, and to trace his own continuous identity as a person across time and space. This is a complex cognitive achievement, which is the establishment of an autobiographical memory system.

Children are required to achieve the capacity to perceive their identity as situated in memory: i.e., 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 terms of autobiography. This is "narrative identity", i.e., a structure that can provide the jumble of autobiographical memories «with some semblance of unity, purpose, and meaning».[44] Research findings show that the complexity and coherence of this structure increase across adolescence until early adulthood.[45]

In this process of narrative self-construction there is an essential psychodynamic component. Dynamic psychology tells us that the affective growth and the construction of identity cannot be separated; the description of the self that since 2-3 years of age the child feverishly pursues is an "accepting description", i.e., a description that is indissolubly cognitive (as *definition* of self) and emotional-affective (as *acceptance* of self). In brief, the child needs a clear and consistent capacity to describe itself, fully legitimized by the caregiver and socially valid. On the other hand, this will continue to hold during the entire cycle of life: the construction of affective life will always be intimately connected to the construction of a well-defined and interpersonally valid identity.[46]

Accordingly, one cannot ascribe concreteness and solidity to one's own self-consciousness if the latter does not possess as a center a description of identity that must be clear and, indissolubly, "good" as worthy of being loved. Our mental balance rests on this feeling of solidly existing as an "I". If the self-description becomes uncertain, the subject soon feels that the feeling of existing vanishes. This can be the result of some psychopathological process; and indeed, clinical research shows that if the coherence of the representation of self is invalidated, or made internally contradictory, then also the primary feeling of self enters into crisis.

Let us consider, e.g., the case of all those patients whose main problem is a chronic feeling of insecurity (or lack of self-esteem, confidence in oneself, solidity of the ego, cohesion of the self – terms that I take to be essentially synonymous). According to a tradition that begins with Michael Balint, Donald Winnicott and John Bowlby, the origin of this "basic fault" is to be traced back mainly to early deficiencies in the relationship between the child and the primary attachment figure.[47] This chronic feeling of insecurity dramatically arises, e.g., in patients with narcissistic personality disorders. A share of narcissistic defenses is normal in the construction of one's identity; pathology comes into play when the subject exploits narcissism to compensate for a condition of insecurity and insufficient self-esteem.

The theme was explored in depth by Heinz Kohut.[48] A narcissistic defense consists not only in the more or less anxious safeguard of the image that we want to have of ourselves, but also in a certain kind of relationship with external world; in this case we talk about an object relation of narcissistic type, viz. a link with situations, things or persons that serve as symbols that help to reassure ourselves about one's identity. In some cases the feeling of identity is so precarious (the self is so little "cohesive", Kohut would say) that the patient finds it difficult to feel existent and is afraid to completely losing contact with himself or herself if deprived of such reassurances.[49]

A collapse of the existential feeling of presence, however, may also occur in cases of sudden breakdown of self-esteem, or unexpected emotional upheavals, or when the continuity of the tissue of our sociality is broken, as can happen when one is suddenly thrown in some dehumanizing total institution.[50] In such circumstances the subject strives to cling to her memories, or to the sense of a projectual dignity, or to the secret security of an affiliation:

> but if all these fail us, then we realize that our mind becomes empty, and not only we no longer know who we are, but also we literally lose the feeling of being present.[51]

Narrative identity, responsibility, guilt

To recapitulate, introspective self-consciousness is an activity of narrative reappropriation of the products of the unconscious information-processing machinery, and this activity has an essentially self-defensive character, being ruled by our primary and universal need to construct and protect a subjective identity whose cohesiveness is the ground of our intra- and interpersonal balances.

But we have to be very clear about one point. When we use cognitive sciences as a source of tools to set up a critique of self-conscious subjectivity, and when we emphasize the defensive nature of self-consciousness as narrative identity, our polestar is a tradition of critical thought that refers to Freud's concept of rationalization. And then it is true that Freud taught us that the description/narration of our inner life gets organized on the basis of a self-apologetic defensiveness, and hence it is a construction permeated by myths and interested self-deceptions. But to this claim the great thinker always associated the firm belief that this self-image can be at least partially "demystified", thus acknowledging the possibility of a path of genuine self-knowledge. Therefore, unlike those trends of thought that cultivate a radically conventionalist view of knowing, the tradition to which Freud belongs draws a clear-cut line of demarcation between the historical truth and the narrative one.[52]

In this framework, the already mentioned contrast between guilt and responsibility gets a new meaning. Locke was definitely right in defining responsibility as the capacity of *critically* re-appropriating one's own acts, projects, memories. But once we have rejected the Lockean theory of inner sense and endorsed an interpretativist view of introspective self-consciousness, the re-appropriation can be defined as "critical" only in the sense of being a self-narration that is more "honest" (less imbued with "bad faith") than that 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.

This is not how things stand for someone who suffers from a sense of guilt: for in this case the subject deceives himself by treating what he feels guilty about extraneous to himself; in short, he expels it from his self-narration.

Let us go back to the driver who, after running over the poor pedestrian, is afflicted by a tormenting sense of guilt and longs for absolution. In his feeling guilty he represents to himself that event as an extraneous body, perceives it as a discontinuity in the flux of his life – in the psychoanalytic idiom, he "evacuates" it. By contrast, if that individual will admit 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, then 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-narration 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 also the crimes or misdemeanors that he has committed.[53]

Notes

[1] This essay is one of a series of papers in which I have been trying to develop a psychodynamic approach to the issue of self-consciousness. See mainly M. MARRAFFA, *Precariousness and Bad Faith. Jervis on the Illusions of Self-Conscious Subjectivity*, in: «Iris. European Journal of Philosophy and Public Debate», vol. III, n. 6, 2011, pp. 171-87; M. MARRAFFA, *Remnants of Psychoanalysis. Rethinking the Psychodynamic Approach to Self-Deception*, in: «Humana.Mente», vol. XX, 2012, pp. 223-243; M. MARRAFFA, E. SIRGIOVANNI, *Coscienza e responsabilità*, in: M. DE CARO, A. LAVAZZA, G. SARTORI (a cura di), *Quanto siamo responsabili? Filosofia, neuroscienze e società*, Codice, Torino 2013, pp. 83-100; M. MARRAFFA, *De Martino, Jervis, and the Self-Defensive Nature of Self-Consciousness*, in: «Paradigmi», vol. XXXI, n. 2, 2013, pp. 109-124.

[2] J. LOCKE, *An Essay Concerning Human Understanding* (1694), Clarendon Press, Oxford 1975, p. 346.

[3] *Ivi*, p. 335.

[4] *Ivi*, p. 345.

[5] See M. DI FRANCESCO, M. MARRAFFA, *The Unconscious, Consciousness, and the Self Illusion*, in: «Dialogues in Philosophy, Mental and Neuro Sciences», vol. VI, n. 1, 2013, pp. 10-22, here p. 11.

[6] M. MARRAFFA, *Introduzione. Giovanni Jervis: la ricerca della concretezza*, in: G. JERVIS, *Contro il sentito dire. Psicoanalisi, psichiatria e politica*, Bollati Boringhieri, Torino 2014, p. LXV-LXVI.

[7] See P.E. GRIFFITHS, *Instinct in the '50s: The British Reception of Konrad Lorenz's Theory of Instinctive Behavior*, in: «Biology and Philosphy»,

vol. XIX, n. 4, 2004, pp. 609-631.

[8] See, e.g., R. HOLT, *Freud Reappraised. A Fresh Look at Psychoanalytic Theory*, Guilford, New York 1989; M. MACMILLAN, Freud Evaluated, MIT Press, Cambridge (MA) 1997, II ed.

[9] See, e.g., E.F. LOFTUS, K. KETCHAM, *The Myth of Repressed Memory*, St. Martin's Press, New York 1994.

[10] See N. MANSON, 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*, Routledge, London 2000, pp. 148-168, here p. 163.

[11] See, e.g., John Bowlby's theory of selective exclusion of information in the chapter 4 of *Attachment and Loss*. Vol. III: *Loss: Sadness and Depression*, Hogarth Press, London 1980.

[12] G. JERVIS, *The Unconscious*, in: M. MARRAFFA, M. DE CARO, F. FERRETTI (eds.), *Cartographies of the Mind*, Springer, Berlin 2007, pp. 147-158, here p. 150.

[13] S. FREUD, *Civilization and Its Discontents* (1930), in: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Hogarth Press and the Institute of Psychoanalysis, London 1961, vol. XXI, pp. 65-66.

[14] *Ibidem*.

[15] See S. MOSCOVICI, *Psychoanalysis: Its Image and Its Public* (1961), Polity Press, Cambridge 2007; R. CASTEL, *Le psychanalysme*, Maspero, Paris 1973.

[16] See M. MARRAFFA, *Introduzione. Giovanni Jervis e la genealogia nascosta della coscienza umana*, in: G. JERVIS, *Il mito dell'interiorità*, Bollati Boringhieri, Torino 2011, pp. XI-LXVI, in particolare p. XXI.

[17] See MARRAFFA, *Remnants of Psychoanalysis*, cit., pp. 226-227.

[18] J. LAPLANCHE, J.-B. PONTALIS, *The Language of Psycho-Analysis* (1967), The Hogarth Press and the Institute of Psycho-Analysis, London 1973, p. 452.

[19] See R.R. HASSIN, J.S. ULEMAN, J.A. BARGH (eds.), *The New Unconscious*, Oxford Umiversity Press, Oxford 2005.

[20] See M. MARRAFFA, A. PATERNOSTER, *Sentirsi esistere. Inconscio, coscienza, autocoscienza*, Laterza, Roma-Bari 2013, pp. 21-22.

[21] D.C. DENNETT, *Consciousness Explained*, Little, Brown and Company, New York 1991, p. 457: «[I]n brains, in computers, in evolution's "recognition" of properties of selected designs».

[22] See M. DI FRANCESCO, M. MARRAFFA, *The Unconscious, Consciousness, and the Self Illusion*, cit., p. 14.

[23] G. JERVIS, *Presenza e identità*, Garzanti, Milano 1984, p. 158.

[24] See J. PIAGET, *The Child's Conception of the World* (1926), Routledge, London 1929; S. MEYER, C. SHORE, *Children's Understanding of Dreams as Mental States*, in: «Dreaming», vol. XI, n. 4, 2001, pp. 179-194.

[25] M. DI FRANCESCO, M. MARRAFFA, *The Unconscious, Consciousness, and the Self Illusion*, cit., p. 15.

[26] See R. NISBETT, T.D. WILSON, *Telling More than we Can Know: Verbal Reports on Mental Processes*, in: «Psychological Review», vol. LXXXIV, n. 3, 1977, pp. 231-259, here p. 233.

[27] E. SCHWITZGEBEL, *Introspection*, in: E.N. ZALTA (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <http://plato.stanford.edu/archives/sum2014/entries/introspection/>.

[28] *Ivi*, §2.1.3.

[29] R. NISBETT, T.D. WILSON, *Telling More than we Can Know*, cit., p. 255.

[30] P. CARRUTHERS, *The Opacity of Mind*, Oxford Umiversity Press, Oxford 2011.

[31] See, e.g., S. DEHAENE, J.-P. CHANGEUX, *Experimental and Theoretical Approaches to Conscious Processing*, in: «Neuron», vol. LXX, n. 2, 2011, pp. 200-227.

[32] See H. FRANKFURT, *Freedom of the Will and the Concept of a Person*, in: «Journal of Philosophy», vol. LXVIII, n. 1, 1971, pp. 5-20; H. FRANKFURT, *The Importance of What We Care About*, Cambridge Umiversity Press, Cambridge 1988.

[33] M. KING, P. CARRUTHERS, *Moral Responsibility and Consciousness*, in: «Journal of Moral Philosophy», vol. IX, n. 2, 2012, pp. 200-228, here pp. 217ff.

[34] See S. POLLO, *L'uso pubblico del naturalismo*, in: «Etica&Politica», vol. IX, n. 2, 2007, pp. 144-147, here p. 146.

[35] P. CARRUTHERS, *The Opacity of Mind*, cit., p.64.

[36] As the advocates of the so-called "Machiavellian intelligence hypothesis" have argued for over 25 years now. See R.W. BYRNE, A. WHITEN, *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Oxford University Press, Oxford 1988.

[37] As it has been suggested in P. RICHERSON, R.

Boyd, *Not By Genes Alone*, University of Chicago Press, Chicago 2005; S. Hrdy, *Mothers and Others*, Harvard University Press, Cambridge (MA) 2009.

[38] See, e.g., S. Nichols, S.P. Stich, *Mindreading*, Oxford Umiversity Press, Oxford 2003.

[39] P. Carruthers, *The Opacity of Mind*, cit., chap. 9.

[40] P. Carruthers, *Mindreading Underlies Metacognition*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 164-176, here p. 167.

[41] See, e.g., J.K. Hamlin, *Failed Attempts to Help and Harm: Intention versus Outcome in Preverbal Infants' Social Evaluations*, in: «Cognition», vol. CXXVIII, n. 3, 2013, pp. 451-474.

[42] On the mechanism of internalization, see D.K. Symons, *Mental State Discourse, Theory of Mind, and the Internalization of Self-Other Understanding*, in: «Developmental Review», vol. XXIV, n. 2, 2004, pp. 159-188.

[43] In regards to this matter, legal language distinguishes between culpable and unintentional antisocial acts, whose damaging effects can be ascribed to the agent but without planning them as such, and malicious acts, where on the contrary there was the (planning) intention to reach that outcome.

[44] D.P. McAdams, B.D. Olson, *Personality Development: Continuity and Change over the Life Course*, in: «Annual Review of Psychology», vol. LXI, 2010, pp. 517-542, here p. 527.

[45] For a review, see R. Fivush, *The Development of Autobiographical Memory*, in: «Annual Review of Psychology», vol. LXII, 2011, pp. 559-582.

[46] See G. Jervis, *Fondamenti di psicologia dinamica*, Feltrinelli, Milano 1993, pp. 317-318.

[47] See M. Balint, *The Basic Fault: Therapeutic Aspects of Regression* (1968), Northwestern Umiversity Press, Evanston (IL) 1992; P. Fonagy, G. Gergely, E.L. Jurist, M. Target, *Affect Regulation, Mentalization and the Development of the Self*, Other Press, New York 2002.

[48] See, e.g., H. Kohut, *The Restoration of the Self*, International Universities Press, New York 1997.

[49] These include what Kohut calls "self-objects", viz. objects of narcissistic type that are experienced as neither internal nor external with respect to the bounds of the identity of person.

[50] See, e.g., the classic E. Goffman, *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*, Doubleday, New York 1961.

[51] G. Jervis, *Il mitodell'interiorità*, cit., pp. 131-132.

[52] According to the interpretive conventionalism, interpretation is 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 verification, tend to disappear, being replaced by a freely creative narrativism of postmodernist type. See, e.g., D.P. Spence, *Narrative Truth and Historical Truth*, W.W. Norton, New York 1982.

[53] See G. Jervis, *Colpa e responsabilità individuale*, in: RAI, «Il Grillo», 6 February 1998, URL = <http://www.emsf.rai.it/grillo/>.