RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2015.0023

ISSN 2039-4667; E-ISSN 2239-2629
Vol. 6 (2015), n. 2, pp. 249-260

Studi

Mindreading and Introspection[1]

Massimo Marraffa

Ricevuto: 5 gennaio 2015; accettato: 16 marzo 2015

Abstract In this article I take a nativist-modularist perspective on mindreading, endorsing the hypothesis that a form of primary mindreading is not a developmental achievement, but an innate social-cognitive evolutionary adaptation implemented by neurocomputational mechanisms that come online during the first year of age. Moreover, I recommend a cognitive-constructivist stance on introspection. Expanding on Peter Carruthers' strong case for the claim that mindreading has a functional and evolutionary priority over introspection, I maintain that mindreading is also developmentally prior to introspection. If the latter is not taken as a competence in isolation, but placed in its context of meaning, i.e., the construction and defense of subjective identity, good reasons emerge for arguing that it takes shape through the act of turning on oneself the capacity to mindread other people; and that this occurs through that socio-communicative interaction with caregivers (and successively other social partners) investigated by the attachment theory.

KEYWORDS: Attachment; Subjective identity; Introspection; Mindreading.

Riassunto *Comprensione della mente altrui e introspezione* – In questo articolo assumo una prospettiva innatistico-modularistica in relazione alla capacità di comprendere le menti altrui (*mindreading*), accogliendo l'ipotesi che una forma primaria di *mindreading* sia non già una conquista ontogenetica bensì un adattamento socio-cognitivo realizzato da meccanismi neurocomputazionali specifici per dominio, già operativi intorno ai 12 mesi di età. Adotterò invece una prospettiva cognitiva e costruttivista sull'introspezione. Estendendo il ragionamento di Peter Carruthers in favore della tesi secondo cui il *mindreading* ha una priorità funzionale e filogenetica sull'introspezione, sosterrò che la prima capacità ha una priorità anche ontogenetica sulla seconda. Se la mentalizzazione in prima persona è presa nel suo contesto di senso, ovvero la costruzione e difesa dell'identità soggettiva, si può sostenere che essa si costituisce nell'atto di rivolgere su se stessi la capacità di mentalizzare in terza persona, e che questo si verifica in virtù di quella interazione socio-comunicativa con il *caregiver* (e successivamente con gli altri partner sociali) che è oggetto di indagine della psicodinamica dell'attaccamento.

PAROLE CHIAVE: Attaccamento; Identità soggettiva; Introspezione; Mindreading.

Introspection I: The self/other parity account

DURING THE 1980S AND 1990S most of the work in Theory of Mind was concerned with the mechanisms that subserve third-person mentalization (henceforth "mindreading"); but in the last decade an increas-

M. Marraffa - Dipartimento di Filosofia, Comunicazione e Spettacolo, Università di Roma Tre,
via Ostiense, 234/236 - 00154 Roma (I)

E-mail: massimo.marraffa@uniroma3.it (✉)

ing number of psychologists and philosophers have also proposed accounts of the mechanisms underlying first-person mentalization (henceforth "introspection"). This required a synergy with other research traditions, most notably the studies on confabulation in cognitive neuropsychology and social psychology.[2]

These research traditions deliver us a huge amount of experiments showing a mismatch between the explanatory *motives* that the subjects report to account for their behavior and the *motivations* (i.e., the multiple real causes) of their behavior. In other words, in these experiments the participants do not have any direct access to the real causes of their behavior; rather, they engage in a rationalization or confabulation, i.e., they make use of socially shared explanatory theories or of an idiosyncratic theorizing, to fabricate reasonable but imaginary explanations of the motivational factors of their behavior.

In this theoretical and experimental framework, the subjects no longer enjoy a privileged access to their own inner life. Rather, they are engaged in an interpretative activity that depends on mechanisms capitalizing on explanatory theories that apply to the same extent to themselves and other people. Such mechanisms are triggered by information about mind-external states of affairs, i.e., the subject's behavior and the situation in which it occurs – information, therefore, in respect to which the subject enjoys no particular epistemic authority. This is a theory of self-knowledge that assumes a "self/other parity".[3]

In social psychology Bem's self-perception theory pioneered a self/other parity account of self-knowledge. With reference to Skinner's methodological guidance, but with a position that reveals affinities with symbolic interactionism, he holds that individuals

> come to "know" their own attitudes, emotions, and other internal states partially by inferring them from observations of their own overt behavior and/or the circumstances in which this behavior occurs.[4]

Nisbett and Wilson developed Bem's approach, claiming that behavioral and contextual data are the input of mechanisms that exploit theories that apply to the same extent to ourselves and to others.[5] In this formulation, the self/other parity account of self-knowledge was welcomed by the theory-theorists in developmental psychology.[6]

It is to be noticed, however, that the self/other parity account is never suggested as an exhaustive theory of self-knowledge; some margin is always left for some sort of direct self-knowledge.[7] Nisbett and Wilson, for instance, 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».[8] 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.[9]

Introspection II: The inner sense account

In order to offer an account of this supposedly direct self-knowledge, some philosophers tried to develop some up-to-date version of the Lockean "inner sense" theory, construing introspection as a process that permits the access to at least some mental phenomena in a relatively direct and non-interpretative way. On this perspective, introspective access does not appeal to theories that serve to interpret behavioral and contextual data, but rather exploits mechanisms that can receive information about inner life through a relatively direct channel.[10]

The attempt to bestow psychological plausibility on the inner sense theory of introspection comes in various forms. Introspection may be realized by a mechanism that processes information about the functional profile of mental states, or their representational content, or both kinds of information.[11] A representationalist-functionalist version of the inner sense theory is Nichols and Stich's account of introspection in terms of monitoring mechanisms.[12] Their hypothesis is that whereas detecting others' mental states and reasoning about one's own and others' mental states are all subserved by the same "Theory of Mind Information", the mechanism for detecting one's own mental states is quite independent of the mechanism that deals with the mental states of other people. More precisely, Nichols and Stich's hypothesis assumes the existence of a set of distinct self-monitoring computational mechanisms, including one for monitoring and providing self-knowledge of one's own perceptual states, and one for monitoring and providing self-knowledge of one's own propositional attitudes.

The monitoring mechanisms account is concerned only with mentalistic self-attribution. As for third-person mentalization and third- and first-person mentalizing reasoning, Nichols and Stich make appeal to the theory-theory. This allows them to restrict the scope of the experiments that show confabulation effects. The errors made by the participants are not about mental-state self-attribution but rather first-person mentalistic reasoning; i.e., understanding the causes of one's own behavior involves reasoning about mental states, and this is definitely a theory-laden process. Thus, if folk-psychological theory is lacking the resources to account for a behavioral sequence, the participant will make inferential errors regarding both one's own inner life and others'. In other terms, self-knowledge can count on two methods: in some circumstances the individuals interpret by exploiting a folk-psychological theory, which may give rise to confabulatory talks; but in other circumstances they can directly and non-interpretatively access one's own mind.[13]

Nichols and Stich see introspection as an inner sense faculty, i.e., a faculty that provides us with a direct quasi-perceptual channel of informational access to our own mental life. This is also Goldman's project, who however tries to relaunch the idea of inner sense within the framework of mental simulation.[14] Here introspection both ontogenetically precedes and grounds mindreading. Mindreaders need to introspectively access their offline products of mental simulation before they can project them onto the target; and this is a form of direct access. Building on Craig's account of interoception, as well as Marr's and Biederman's computational models of visual object recognition, Goldman maintains that introspection is a perception-like process that involves a transduction mechanism that takes neural properties of mental states as input and outputs representations in a proprietary code, which code represents types of mental categories and classifies mental-state tokens in terms of such categories.[15]

To recapitulate. In this section we have discussed the approach of some philosophers who acknowledge the theoretical, and hence non-introspective, character of first-person knowledge of the causes of our thoughts and behavior, and nevertheless continue to think that in some specific cases the access to one's mental life is direct and non-interpretative. Nichols and Stich's theory of introspection postulates mechanisms that are fed, through a relatively direct channel, by information about perceptual and propositional attitude states. Goldman argues that the mindreader needs to introspectively access its offline products of mental simulation before it can project them onto the target; and introspection is a perception-like process. As we will see in the next section, however, both theories are vulnerable to Carruthers' criticism of the idea of a non-interpretative access to propositional attitudes.

Introspection III: Self-interpretation plus sensory access

In opposition to the attempt to develop a cognitively plausible inner sense view of introspection (both in Nichols and Stich's version as well as in Goldman's), Carruthers developed a very sophisticated version of the self/other parity account: the interpretive sensory-access (ISA) account of the nature and sources of self-knowledge.[16] According to the ISA account, although we can have non-interpretive access to our own sensory and affective states, the self-attribution of propositional attitude states is always, in agreement with the self/other parity account, a swift and unconscious process of self-interpretation that exploits the same sensory channels that we utilize when working out other people's mental states.

In order to account for the conscious accessibility of our perceptual states, the ISA account assumes the validity of the global workspace models of human neurocognitive architecture. There is now extensive evidence supporting such models;[17] moreover, analyses of functional connectivity patterns in the human brain have demonstrated just the sort of neural architecture necessary to realize the main elements of a global broadcasting account.[18]

More specifically, these studies show the existence of two main neurocomputational spaces within the brain, each characterized by a distinct pattern of connectivity.[19] The first space is a processing network, composed of a set of parallel, distributed, and functionally specialized processors or modular subsystems characterized by highly specific local or medium-range connections. The subsystems compete each other to access the global neuronal workspace, which is implemented by long-range cortico-cortical connections, mostly originating from the pyramidal cells of layers 2 and 3 that are particularly dense in prefrontal, parieto-temporal and cingulate associative cortices, together with their thalamocortical loops. When one of these subsystems accesses the global neuronal workspace, its outputs (i.e., *sensory* information including perceptions of the world, the deliverances of somatosensory systems, imagery and inner-speech) are broadcast to an array of concept-using consumer systems – e.g., systems that use the perceptual input to form judgments or make decisions.[20]

Among the conceptual judgment-forming systems there is a mindreading system which, drawing on a folk-psychological theoretical framework, generates metarepresentational beliefs about the mental states of others and of oneself. This system 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 what about the outputs of the other consumer systems, i.e., propositional-attitude events?

Most philosophers assume that propositional attitudes are consciously accessible, relying on

> a conception of the mind as containing, at its core, a workspace in which thoughts can be created, reflected on, and evaluated, and in which attitudes of all types can be active and enter into processes of reasoning and thinking.[21]

However, Carruthers replies, the only central workspace in the human mind is the *working memory system*, which utilizes the mechanisms of global broadcast to subserve a wide variety of central-cognitive purposes. And what can be found within working memory are not propositional attitudes, but rather imagery, inner speech, and so forth; working memory's operations are always *sensory based*. Indeed, there are good reasons for thinking that propositional attitudes are not capable of being globally broadcast, and hence can never be consciously accessible.[22]

Since there are also good reasons for thinking that there are no causal pathways from the outputs of the consumer systems to mindreading system,[23] the latter must exploit the globally broadcast perceptual information, together with some forms of stored knowledge, to infer the agent's propositional attitudes, precisely as it happens with third-person mindreading. Thus, as already mentioned, self-attribution of propositional attitudes always occurs by means of a process of *self-interpretation*, which rests on the sensory awareness of data concerning one's own behavior, contextual data and/or sensory items in working memory.

In addition to experimental findings about the nature and sensory basis of broadcasting and working memory, Carruthers defends his ISA account taking position on (i) the nature and source of our capacities for metacognitive control of learning and reasoning; (ii) alleged dissociations between self-knowledge and other-knowledge in autism and schizophrenia; and (iii) the above-mentioned studies on confabulation in cognitive neuropsychology and social psychology. Let us quickly consider (i)-(iii).

Considerations concerning the evolutionary role of mindreading and the literature on metacognition

The ISA account posits a single phylogenetic route for both mindreading and introspection – an integrated faculty of metarepresentation evolved for mindreading and later exapted for introspection. This is what is legitimate to expect in light of the hypothesis that mindreading, as an ingredient essential to our social intelligence, evolved to provide an adaptive advantage in pursuing the aims of two motivational macrosystems, the first committed to self-assertiveness and competition; the second aimed to prosociality and cooperation.[24]

It has been objected that meta-representational mindreading is likely to be a late «exaptation derived from linguistic abilities and general-purpose concept learning resources».[25] To this Carruthers replied that there are at least two problems with this view.

First, a series of investigations using nonverbal, spontaneous-response versions of false-belief tasks provides evidence that metarepresentational mindreading is already present in infants around the middle of the first year of life, implemented by the domain-specific component of the mindreading system.[26] Second, metarepresentation is required for lexical acquisition; if children were not able to grasp the speaker's referential intentions, learning the meanings of words would not be possible.[27]

In contrast with the ISA account, inner sense theories bear an explanatory burden. For if introspection and mindreading are implemented by two (or more) neurocognitive mechanisms, then a distinct account of the evolution of each is needed. One hypothesis about the kind of evolutionary pressure that can account for the emergence of first-person mentalizing mechanism(s) is that the capacity to represent one's own mental states (or some subset thereof) evolved first, presumably to enable organisms to increase the advantages of metacognitive monitoring and control.[28] Once evolved, the conceptual and inferential resources involved were somehow exapted for mindreading.[29]

However, the hypothesis that introspection evolved for metacognitive purposes does not tally with the available evidence. The human and comparative metacognitive data seem to show at least two things. First, in many cases the controlling function of metacognition does not involve any introspective capacity.[30] Second, our metacognitive interventions are not capable of the sort of direct impact on cognitive processing that would be predicted if metacognition had, indeed, evolved for the purpose.[31]

Dissociation data. Since Nichols and Stich's monitoring mechanisms account assumes that introspection does not involve mechanisms of the sort that figure in mindreading, it implies that the first capacity should be dissociable from the second. Accordingly, they make the hypothesis of a double dissociation between schizophrenia and autism. In adults with Asperger's syndrome the capacity of detecting their own mental states would be intact despite of the mindreading deficit; the opposite pattern would be observed in schizophrenic patients with passivity experiences.[32]

The ISA account predicts that this dissociation should not occur, since there is just a single faculty involved in both mindreading and introspection. Consequently, Carruthers recruits data that refute Nichols and Stich's hypothesis. For example, Williams and Happé showed that in children with autism spectrum disorder the capacity to attribute intentions to themselves is just as impaired as is the capacity to attribute intentions to others, and that both poor performances can be imputed to the difficulties that ASD children have with mindreading in general.[33] With regard to schizophrenia, Carruthers points out that passivity experiences are not best explained by the impairment of a system subserving first-person mindreading. A more fruitful hypothesis is the failure of the "comparator system".[34]

Confabulation data. Even more than phylogenetic and psychopathological considerations, the central prediction made by the ISA account is *frequent confabulation*, which serves to distinguish it empirically from the inner-sense theories of self-knowledge.[35] As seen above, the inner-sense theorists accommodate the confabulation data by postulating two methods – there would be an introspective but also an interpretive route to our own attitudes. Consequently, inner-sense theories are less simple than the ISA account. Still more important, unlike inner-sense theorists' dual-method hypothesis, the ISA account can explain the overall patterning of the confabulation data. Since Carruthers' model holds that the knowledge of one's propositional attitudes rests on a theory-driven interpretive process that is fed by sensory and behavioral data, it predicts confabulation effects anytime such data are misleading, or the theories that the subjects use to interpret themselves are inadequate. The dual-method theorists, instead, will be in trouble to provide some principled account of the circumstances in which people access their propositional attitudes directly and the circumstances in which they rely on self-directed mindreading.

The developmental asymmetry between introspection and mindreading

Carruthers' ISA account holds that mindreading has a functional and evolutionary priority over introspection, but it does not predict that the former is *developmentally* prior to the latter.[36]

However, Carruthers also advances the hypothesis that the mindreading system must contain a model of what minds are and of «the access that agents have to their own mental states».[37] Such a model is likely to be essentially "Cartesian", assuming that subjects know, immediately and without self-interpretation, what they are experiencing, judging and intending. This assumption, Carruthers speculates, may have great heuristic value, greatly simplifying the mindreading system's computations. But then he also notices that an alternative account to his is outlined by T.D. Wilson, who suggests that the self-transparency assumption

> may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image. In fact, *both* accounts might be true.[38]

Moreover, Carruthers contemplates Wilson's hypothesis again, holding that the claim that the emergence of introspection is a by-product of the evolution of mindreading is compatible with the hypothesis that the former

> might have come under secondary selection thereafter, perhaps by virtue of helping to build and maintain a positive self-image.[39]

Thus, here Carruthers is opening the door to the psychodynamic topic of defense mechanisms, i.e., the hypothesis that our activity of re-appropriation of the products of the neurocomputational unconscious is ruled by a self-apologetic defensiveness.

On the other hand, the above-mentioned studies in social psychology which investigated how human behavior can respond to motivational factors that are not available to introspection and verbal report, as well as the extended literature on causal attributions, have their main origin in Freud and psychoanalysis. There is a problem, though. Carruthers' focus is *not* on self-knowledge construed as «awareness of oneself as an ongoing bearer of mental states and dispositions, who has both a past and a future».[40] His focus – he makes clear – is *knowledge of one's own current mental states*; and this knowledge «is arguably more fundamental than knowledge of oneself as a self with an ongoing mental life».[41]

Now, insofar as Carruthers takes introspection *merely* as a competence to self-attribute one's own current mental states, Wilson's hypothesis of the self-defensive nature of introspection cannot be built into the ISA theory. As it will be made clear below, the psychodynamic topic of defenses makes sense only in the context of the construction and protection of the psychological (as opposed to bodily) self-consciousness, or subjective identity. However, once introspection is placed into this context, it becomes possible to argue that it develops through the act of turning on oneself the competence to mind-read others; and that this occurs through that socio-communicative interaction with caregivers which is the subject matter of the psychodynamics of attachment.

These different ways of viewing introspection seem to be what is at stake in an exchange between Fernyhough and Carruthers on BBS's pages.[42] Fernyhough draws the attention to some sources of evidence for the hypothesis of a late emergence of the child's inner experience – in particular, the findings that the transformation of private speech into inner speech may not be complete until middle childhood, and that visual imagery also takes time to develop.[43]

Since inner speech and visual imagery are among the data that feed the interpretive process underlying the knowledge of one's propositional attitude states, Fernyhough concludes that the emergence of introspection should be developmentally constrained by the emergence of inner speech and visual imagery.

Given what we know about the timetable for the emergence of mindreading abilities (especially the already mentioned evidence for very early mindreading competences), Carruthers' theory should predict a developmental lag between mindreading and introspection. However, Carruthers denied that there is such implication.[44]

The problem here seems to be that Carruthers and Fernyhough are approaching introspection from very different perspectives. As said, the former's focus is on a minimal sense of introspection as competence to self-attribute one's own current mental states taken independently from any cognition of oneself as a self construed as introspective self-description, i.e., the psychological self-consciousness or subjective identity.

By contrast, Fernyhough's focus is precisely on the development of introspective self-consciousness in a Vygotskian perspective – an outward-in construction that occurs in an interpersonal context, namely in the relationship with caregivers and peers. Thus, Carruthers takes introspection as a competence *in isolation*, and this notion is «too restrictive» to elaborate our understanding of its development beyond «the standard strategy of comparing children's performance across false-belief tasks».[45]

Fernyhough, in contrast, sets introspection back in its context of meaning, one in which the turning of one's mindreading abilities upon oneself is seen as part of the construction of an inner experiential space, and then of an autobiographical self. It is introspection taken in this constructive dimension that is relevant to the psychodynamic topic of defenses. In order to make this point clear we have to turn our attention to the relation between mindreading and attachment theory.

Attachment, mentalization, and psychopathology

According to the psychodynamics of attachment, the primordial psychological need of the very young child, around which his mental life gradually takes shape, pertains to the physical contact and the construction of protective and communicative interpersonal structures. In this perspective, several attachment theorists and developmental psychologists have put forward different versions of the hypothesis that there is a direct ontogenetic causal and functional link between security of infant attachment and its early interactive predictors on the one hand, and the development of explicit mindreading abilities in later childhood on the other.

However, evidence does not confirm the hypothesis that the security of the attachment relationship is *directly* related to children's mentalization development. According to Meins, the observed link between attachment security and mentalization is *indirect*, with both attachment security and mentalization performance being predicted by caregivers' *mind-mindedness*, i.e., the proclivity to treat one's infant as an individual with a mind, rather than merely an entity with needs that must be satisfied. In particular, an aspect of the caregivers' internal-state language, namely comments that appear to be appropriate to the mental state of the child, is a predictor of children's understanding of mind.[46]

Now, there is no doubt that caregiver-child communicative interaction impacts on the development of mentalization; the problem is *how* the child's exposure to such interaction can have such an impact. In particular, the role that language plays in this context needs to be clarified. Jill De Villiers, for example, would disagree with Meins' hypothesis that language, in the form of comments that appear to be appropriate to the mental state of the child, is crucial as element that is able to impact on the development of mentalization. More radically, De Villiers thinks that our mentalistic abilities are *constituted* by language; more specifically, mastery of the grammatical rules for embedding tensed complement clauses under verbs of speech or cognition provides children with a necessary representational format for dealing with false beliefs.[47]

It has been shown, however, that correlation between linguistic exposure and mindreading does not depend on the use of specific grammatical structures; syntax is not constitutive of the mentalizing abilities of adults; and mastery of sentence complements is not even a necessary condition of the development of mindreading in children.[48] But above all, any theorizing on the relation between language and mentalization must come to grips with the already-cited investigations that, using "spontaneous-response" tasks, seem to show that metarepresentational abilities emerge from a specialized neurocognitive mechanism that matures during the second year of life. Such evidence knocks out a constitution-thesis *à la* De Villiers, but also raises a problem for Meins' hypothesis, which should be more prudently construed in terms of a form of scaffolding that initially is not linguistic.[49]

Thus, there seems to be no direct ontogenetic causal and functional link between the quality of early infant attachment – or the linguistic scaffolding consisting in mothers' internal-state talk that is appropriately attuned to the infant's thoughts and feelings – on the one hand, and the development of mindreading on the other. The theory of attachment builds within a contextualist and systemic framework, where (individual) biology and (social) relationality cannot be separated. Individuals are pre-wired to the interpersonal relationship from the birth, and mindreading is part and parcel of such pre-organization. Therefore, our competence to mindread others is not a developmental achievement, but

> an innate social-cognitive evolutionary adaptation implemented by a specialized and pre-wired mindreading mechanism that seems active and functional at least as early as 12 months of age in humans.[50]

An adaptation, therefore, independent of the attachment system. When we take into consideration introspection, in contrast, the relationship between attachment and mentalization is no longer simply a "scaffolding" one: the child's socio-communicative interaction with caregivers is *constitutively involved* *in* the construction of the virtual inner space of the mind, or introspective self-consciousness. The approach to first-person mentalization is then less "neurocognitively guaranteed" and more markedly constructivist compared with the mindreading abilities.

In this perspective, introspective self-consciousness takes shape in the child in a relationship with caregivers that is made of words, descriptions, designations, evaluations of the person. Through the dialogue with caregivers (and then with other social partners) children construct their own identity, both objective (for others) and subjective (for itself). And in the perspective of symbolic interactionism, the identity-for-itself can be said to arise out of the identity-for-others; introspective self-description takes shape through a process of internalization of the ways in which others see and define us. As Gergely puts it,

> the intentional actions and attitudes repeatedly expressed towards the young child by caregivers and peers serve as the inferential basis for attributing generalized intentional properties to the self in an attempt to rationalize the social partners' self-directed behavior.[51]

The development of introspection is, then, the process through which a subject constructs itself as psychologically self-conscious (and not only as physically self-conscious) in an interplay of mindreading, autobiographical memory, and socio-communicative capacities modulated by socio-cultural variables.

The young child who turned his mindreading abilities upon himself under the thrust of caregivers' mind-minded talk, by the end of the preschool years begins to grasp his introspective self-description as rationalized in terms of autobiography.[52] This is self-knowledge in its most demanding form, requiring

> a conception of oneself as a self, together with a capacity for narrative, weaving one's current thoughts and experiences into a larger story of one's life.[53]

We are now in condition to see the connection between introspective self-description and the psychodynamic topic of defenses. Breaking with a long philosophical tradition that has viewed self-consciousness as a purely cognitive phenomenon,[54] the psychodynamics of attachment teaches us that affective growth and construction of identity cannot be separated. The description of the self that from 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-affectional (as *acceptance* of self). Briefly, 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 be the case during the entire cycle of life: 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 presence fades.[55] It is here, therefore, with regard to the construction and defense of a well-defined and interpersonally valid identity, that the idea of a direct ontogenetic causal and functional link between attachment and (first-person) mentalization finds its application.

Conclusions

In this article I have taken a nativist-modularist perspective on mindreading, endorsing the hypothesis that a form of primary mindreading is not a developmental achievement, but an innate social-cognitive evolutionary adaptation implemented by neurocomputational mechanisms that are active and functional by the first year of age.

Moreover, I adopted a cognitive-constructivist stance on introspection. Expanding on Carruthers' strong case for the claim that mindreading has a functional and evolutionary priority over introspection, I maintained that mindreading is also developmentally prior to introspection. If the latter is not taken as a competence in isolation, but placed in the context of the construction and defense of subjective identity, good reasons emerge for arguing that it develops through the act of turning on oneself the capacity to mindread other people; and that this occurs through that socio-communicative interaction with caregivers (and successively other social partners) investigated by the attachment theory.

Notes

[1] This article builds on two previous papers: M. MARRAFFA, *The Unconscious, Self-consciousness, and Responsibility*, in: «Rivista Internazionale di Filosofia e Psicologia», vol. V, n. 2, 2014, pp. 207-220; M. MARRAFFA, C. MEINI, *La priorità della mentalizzazione in terza persona: implicazioni per la teoria dell'attaccamento*, in: «Attaccamento e sistemi complessi», vol. II, n. 1, 2015, pp. 45-64. I am very grateful to Cristina Meini for comments on a previous version of this article.

[2] 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; D. WEGNER, *The Illusion of Conscious Will*, MIT Press, Cambridge (MA) 2002; T. WILSON, *Strangers to Ourselves*, Harvard University Press, Cambridge (MA) 2002.

[3] E. SCHWITZGEBEL, *Introspection*, in: E.N. ZALTA (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <http://plato.stanford.edu/archives/sum 2014/entries/introspection/>, §2.1.

[4] D.J. BEM, *Self-Perception Theory*, in: L. BERKOWITZ (ed.), *Advances in Experimental Social Psychology*, Academic Press, New York 1972, vol. VI, pp. 1-62, here p. 5. See also G. Ryle's well-known passage: «The sort of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same […] in principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe» (G. RYLE, *The Concept of Mind* (1949), Routledge, London 2009, p. 139).

[5] This is well illustrated by the studies based on actor-observer comparisons. See R. NISBETT, T.D. WILSON, *Telling More Than we can Know*, cit., pp. 250-251; T.D. WILSON, J.I. STONE, *Limitations of Self-knowledge: More on Telling More Than we can Know*, in: P. SHAVER (ed.), *Review of Personality and Social Psychology*, Sage, Beverly Hills (CA) 1985, vol. VI, pp. 167-183.

[6] See A. GOPNIK, *How we Read our Own Minds: The Illusion of First-person Knowledge of Intentionality*, in: «Behavioral and Brain Sciences», vol. XVI, n. 1, 1993, pp. 1-14.

[7] E. SCHWITZGEBEL, *Introspection*, cit., §2.1.3.

[8] R. NISBETT, T.D. WILSON, *Telling More Than we can Know*, cit., p. 255.

[9] E. Schwitzgebel gives textual evidence that also Bem, Gopnik and Ryle are prone to this criticism.

[10] "Inner sense" theories of introspection in P. CARRUTHERS, *The Opacity of Mind*, Oxford University Press, Oxford 2011, chap. 7. "Inside access" view of introspection in P. ROBBINS, *The Ins and Outs of Introspection*, in: «Philosophy Compass», vol. I, n. 6, 2006, pp. 617-630, here p. 618. "Self-detection" accounts of self-knowledge in E. SCHWITZGEBEL, *Introspection*, cit., §2.2.

[11] P. ROBBINS, *The Ins and Outs of Introspection*, cit., pp. 618-619.

[12] S. NICHOLS, S. STICH, *Mindreading*, Oxford University Press, Oxford 2003.

[13] A. GOLDMAN, *Simulating Minds*, Oxford University Press, Oxford 2006, p. 232.

[14] *Ivi*, pp. 258-275.

[15] In the same vein, the hypothesis has been put forward that we mentally induce the internal states of the other in ourselves through "neuronal resonance". See V. GALLESE, C. KEYSERS, G. RIZZOLATTI, *A Unifying View of the Basis of Social Cognition*, in: «Trends in Cognitive Sciences», vol. VIII, n. 9, 2004, pp. 396-403.

[16] P. CARRUTHERS, *The Opacity of Mind*, cit.; P. CARRUTHERS, *Mindreading the Self*, in: S. BARON-COHEN, H. TAGER-FLUSBERG, M. LOMBARDO (eds.), *Understanding Other Minds: Perspectives from Social Cognitive Neuroscience*, Oxford University Press, Oxford 2013, pp. 467-485.

[17] 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.

[18] See, e.g., E. BULLMORE, O. SPORNS, *Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems*, in: «Nature Reviews Neuroscience», vol. X, n. 4, 2009, pp. 186-198; M. SHANAHAN, *Embodiment and the Inner Life*, Oxford University Press, Oxford 2010.

[19] See, e.g., S. DEHAENE, J.-P. CHANGEUX, L. NACCACHE, *The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications*, in: S. DEHAENE, Y. CHRISTEN (eds.), *Characterizing Consciousness: From Cognition to the Clinic?*, Springer, Berlin Heidelberg 2011, pp. 55-84.

[20] Judgements are «events of belief formation»; decisions are «acts of willing, or the events that create novel activated intentions» (see P. CARRUTHERS, *Introspection: Divided and Partly Eliminated*, in: «Philosophy and Phenomenological Research», vol. LXXX, n. 1, 2010, pp. 76-111, here p. 78).

[21] P. CARRUTHERS, *On Central Cognition*, in: «Philosophical Studies», vol. CLXX, n. 1, 2014, pp. 143-162, here p. 145.

[22] With two exceptions: sensorily-embedded judgments (seeing as, hearing as, and so on) and affective feelings directed toward some object or situation are propositional attitudes that may figure in the central workspace.

[23] P. CARRUTHERS, *The Opacity of Mind*, cit., §1.3.

[24] See J.D. LICHTENBERG, *Psychoanalysis and Motivation*, Analytic Press, Hillsdale (NJ) 1989; G. JERVIS, *Psicologia dinamica*, Il Mulino, Bologna 2001.

[25] C. BUCKNER, A. SHRIVER, S. CROWLEY, C. ALLEN, *How "Weak" Mindreaders Inherited the Earth*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 140-141, here p. 140.

[26] P. CARRUTHERS, *Mindreading in Infancy*, in: «Mind & Language», vol. XXVIII, n. 2, 2013, pp. 141-172.

[27] P. CARRUTHERS, *Mindreading Underlies Metacognition*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 164-176, here p. 166.

[28] See, e.g., J. COUCHMAN, M. COUTINHO, M. BERAN, D. SMITH, *Metacognition is Prior*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, p. 142.

[29] This could have happened in two ways: either these first-person resources were redeployed to form the basis of a distinct mentalization faculty of the sort defended by Nichols and Stich; or they were combined with emerging capacities for imaginative perspective-taking to enable simulations of the mental lives of others, as Goldman suggests.

[30] One example is the "comparator system", one of the main components of the action-control system, which does not involve any metarepresentations. See P. CARRUTHERS, *How we Know our Own Minds: The Relationship Between Mindreading and Metacognition*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 121-138, here p. 135.

[31] P. CARRUTHERS, B. RITCHIE, *The Emergence of Metacognition: Affect and Uncertainty in Animals*, in: M. BERAN, J. BRANDL, J. PERNER, J. PROUST (eds.), *Foundations of Metacognition*, Oxford University Press, Oxford 2012, pp. 76-93.

[32] See S. NICHOLS, S. STICH, *Mindreading*, cit., chap. 4.

[33] D. WILLIAMS, F. HAPPÉ, *Representing Intentions in Self and Other: Studies of Autism and Typical Development*, in: «Developmental Science», vol. XIII, n. 2, 2010, pp. 307-319.

[34] See above, note 30. On this hypothesis C. FRITH, S-J. BLAKEMORE, D. WOLPERT in *Explaining the Symptoms of Schizophrenia: Abnormalities in the Awareness of Action*, in: «Brain Research Reviews», vol. XXXI, n. 2, 2000, pp. 357-363.

[35] P. CARRUTHERS, *The Opacity of Mind*, cit., pp. 6, 365.

[36] P. CARRUTHERS, *Mindreading Underlies Metacognition*, cit., p. 167.

[37] P. CARRUTHERS, *How we Know Our Own Minds*, cit., p. 127.

[38] P. CARRUTHERS, *Cartesian Epistemology: Is the Theory of the Self-transparent Mind Innate?*, in: «Journal of Consciousness Studies», vol. XV, n. 4, 2008, pp. 28-53, here p. 138, note 42. Wilson's hypothesis is in *Strangers to Ourselves*, cit. above, note 12.

[39] P. CARRUTHERS, *How we Know Our Own Minds*, cit., here p. 128.

[40] P. CARRUTHERS, L. FLETCHER, B. RITCHIE, *The Evolution of Self-knowledge*, in: «Philosophical Topics», vol. XV, n. 2, 2012, pp. 13-37, here p. 14.

[41] *Ibidem*.

[42] C. FERNYHOUGH, *What can we Say About the Inner Experience of the Young Child?*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 143-144; P. CARRUTHERS, *Mindreading Underlies Metacognition*, cit., p. 167.

[43] See, e.g., A.S. AL-NAMLAH, C. FERNYHOUGH, E. MEINS, *Sociocultural Influences on the Development of Verbal Mediation: Private Speech and Phonological Recoding in Saudi Arabian and British Samples*, in: «Developmental Psychology», vol. XLII, n. 1, 2006, pp. 117-131; C. FERNYHOUGH, K.A. BLAND, E. MEINS, M. COLTHEART, *Imaginary Companions and Young Children's Responses to Ambiguous Auditory Stimuli: Implications for Typical and Atypical Development*, in: «Journal of Child Psychology and Psychiatry», vol. XLVIII, n. 11, 2007, pp. 1094-1101.

[44] All that follows, Carruthers writes, «is that there will be many more moments in the daily lives of children at which they will be unwilling to

attribute occurrent thoughts to themselves than is true of the daily lives of adults, because the conscious mental events that might underlie such self-attributions simply are not present. Nothing follows about children's competence to self-attribute attitudes. Nor does it follow that children will be weaker at attributing attitudes to themselves than they are at attributing attitudes to others, provided that the tasks are suitably matched» (see P. CARRUTHERS, *Mindreading Underlies Metacognition,* cit., p. 167).

[45] See M. HERNIK, P. FEARON, P. FONAGY, *There Must be More to Development of Mindreading and Metacognition Than Passing False Belief Tasks*, in: «Behavioral and Brain Sciences», vol. XXXII, n. 2, 2009, pp. 147-148, here p. 147.

[46] E. MEINS, *Social Relationships and Children's Understanding of Mind: Attachment, Internal States, and Mind-mindedness*, in: M. SIEGAL, L. SURIAN (eds.), *Access to Language and Cognitive Development*, Oxford University Press, Oxford 2011, pp. 23-43.

[47] J. DE VILLIERS, *Can Language Acquisition Give Children a Point of View?*, in: J.W. ASTINGTON, J.A. BAIRD (eds.), *Why Language Matters for Theory of Mind*, Oxford University Press, Oxford 2005, pp. 186-219.

[48] See P. CARRUTHERS, *Language in Cognition*, in: E. MARGOLIS, R. SAMUELS, S. STICH (eds.), *The Oxford Handbook of Philosophy of Cognitive Science*, Oxford University Press, Oxford 2011, pp. 382-401, here pp. 389-390.

[49] M. MARRAFFA, C. MEINI, *La priorità della mentalizzazione in terza persona*, cit., .

[50] G. GERGELY, Z. UNOKA, *Attachment, Affect-regulation and Mentalization*, in: E.L. JURIST, A. SLADE, S. BERGNER (eds.), *Mind to Mind*, Other Press, New York 2008, pp. 50-87, here p. 58.

[51] G. GERGELY, *The Development of Understanding Self and Agency*, in: U. GOSWAMI (ed.), *Blackwell Handbook of Childhood Cognitive Development*, Blackwell, Oxford 2002, pp. 26-46, here p. 42.

[52] See R. FIVUSH, *The Development of Autobiographical Memory*, in: «Annual Review of Psychology», vol. LXII, n. 2, 2011, pp. 570-571.

[53] P. CARRUTHERS, L. FLETCHER, B. RITCHIE, *The Evolution of Self-knowledge*, cit., p. 14.

[54] See, e.g., J. BERMÚDEZ, *Self-Consciousness*, in: M. VELMANS, S. SCHNEIDER (eds.), *The Blackwell Companion to Consciousness*, Blackwell, Oxford 2007, pp. 456-467, here p. 456: «Self-consciousness is primarily a cognitive, rather than an affective state».

[55] G. JERVIS, *Presenza e identità*, Garzanti, Milano 1984, p. 50.