Chapter 11

Attachment, affect-regulation and mentalization:

The developmental origins of the representational affective self

György Gergely and Zsolt Unoka

In this chapter we propose an evolutionary-based social-cognitive theory of the early
development of the representational affective self in humans and its role in emotional self-
regulation and control. We first identify species-unique properties of human caregiver-infant
interactions and critically discuss alternative proposals concerning the functional role of the
human attachment system in the development of mentalizing on the one hand, and emotional
self-awareness and affective self-control on the other. We propose that the human-specific
features of early caregiver-infant interactions provide the necessary input conditions for
specialized representation-building and attention socialization mechanisms (such as contingency
detection, social biofeedback, and natural ‘pedagogy’). These mechanisms play a crucial role in
establishing primary self-other affective relationship representations that capture the
characteristic causal structure of contingent reactivity of early attachment relationships. They
also set up cognitively accessible second-order emotion representations with associative links to
the self’s procedurally represented pre-wired basic emotions. We then characterize different
levels of primary and secondary emotion-control systems and show how the developmentally
established affective self-other representations sub-serve these emotion-regulative mechanisms.
Finally, we argue that the socially constructed second-order emotion representations support the
ontogenetic extension of the domain of mentalizing to include the self’s own causal mental states
(apart from those of others) and show how this enables the functional use of mentalization for the purposes of emotional self-regulation and control in affectively charged interactions and relationships in humans.

11.1 Species-unique properties of human attachment interactions and their functions: Challenging current views on the development of mentalizing

Humans are just one of the many social animal species with an attachment instinct system where the quality of early caregiving (Bowlby 1969; Harlow 1961; Hofer 1995; Polan and Hofer 1999; Suomi 1995, 1999) has formative effects on the organism’s later capacity to adaptively function in affiliative relationships. In contrast, humans may be unique in their ability to rely on a specialized inferential and representational system for mentalizing (or ’theory-of-mind’, see Baron-Cohen 1995; Baron-Cohen et al. 1985; Dennett 1978, 1987; Fodor 1992; Leslie 2000; Perner 1991; Wellman 1990) that has evolved to predict, interpret, and manipulate others’ actions through inferring and attributing causal intentional mind states (such as desires, intentions, and beliefs) to them.

Arguably, attachment and mentalizing are two independent adaptations selected to serve qualitatively different evolutionary functions. This is suggested, for example, by the fact that many social species (e.g., rodents or rhesus monkeys, see Hofer, 1995; Polan and Hofer, 1999; Suomi, 1999) that apparently lack the capacity for mentalizing, nevertheless, possess an innate infant-caregiver attachment system. According to Bowlby’s (1969) original proposal, the basic evolutionary function of the attachment instinct is to provide a specialized interactive mechanism to ensure predator evasion for the immature offspring through a pre-wired goal-directed signal system serving proximity seeking and maintenance between caregiver and infant. This function
certainly seems independent and qualitatively different from the primary function served by the specialized social-cognitive adaptation for mentalizing, which is to predict and interpret the actions of conspecifics during adult competitive or cooperative situations through attributing causal mental states to them.

It is noteworthy, however, that the structural organization of early attachment interactions in humans has become qualitatively enriched by species-unique evolved features that are conspicuously absent from mother-infant attachment interactions of other social species (including our closest primate relatives). Thus, human mother-infant interactions show a unique “proto-conversational” turn-taking contingency structure (Brazelton and Tronick 1980; Brazelton et al. 1974; Sander 1988; Stern 1985; Trevarthen 1979; Trevarthen and Aitken 2001; Tronick 1989; Tronick and Cohn 1989). Human infants exhibit early sensitivity and preference for highly response-contingent stimulus events characteristic of the interactive style of infant-attuned social partners (Bigelow 1999; Bigelow and De Coste 2003; Bigelow and Rochat 2006; Watson 1972, 1985, 1994; Bahrick and Watson 1985; Rochat and Morgan 1995; Lewis et al. 1990). They also show special sensitivity to so-called ostensive-communicative cues (Csibra and Gergely 2006) such as eye-contact (Farroni et al. 2002), infant-directed speech or ’motherese’ (Fernald 1985, 1992; Cooper and Aslin 1990), and contingent reactivity (Floccia et al. 1997; Johnson et al. 1998; Movellan and Watson 2002; Watson 1994, 1995, 2001). Young human infants spontaneously attend to and gaze-follow others’ referential cues (such as gaze shift), but only if these are preceded by direct eye-contact or infant-contingent reactivity (Farroni et al. 2002; Johnson et al. 1998; Movellan and Watson 2002; see Csibra and Gergely 2006, for a review). Unlike apes, by 12 months humans show a communicative and referential understanding and use of pointing in triadic joint attention interactions (Behne et al. 2005;
Liszkowski *et al.* 2004; Tomasello *et al.* in press; Southgate *et al.* in press). Human mother-infant interactions also involve frequent *exchanges of an increasingly differentiated repertoire of basic emotion expressions* (Camras 1992; Gergely 2002, 2007a; Gergely and Watson 1996, 1999; Sroufe 1996; Tronick 1989; Cohn and Tronick 1988; Izard and Malatesta 1987; Bennett *et al.* 2004, 2005; Malatesta *et al.* 1989). This contrasts with the small range and low frequency of emotion displays in non-human infant-caregiver attachment interactions. Even in primates mother-directed affect displays by infants are rare and restricted to expressions of distress and protest triggered by loss of proximity (e.g., De Marco and Visalberghi 2007). Empathic *affect-mirroring emotion displays* by the caregiver are also unique to human mother-infant interactions (Bigelow 1999; Gergely 2004, 2007a; Gergely and Watson 1996, 1999; Tronick 1989; Cohn and Tronick 1988; Fonagy *et al.* 2002 in press; Malatesta *et al.* 1989; Uzgiris *et al.* 1989). In this chapter we shall explore the question whether these species-unique properties of human attachment interactions are related functionally or developmentally to the emergence of the capacity to mentalize or whether they serve other basic functions that are independent or at least not directly related to mindreading.

In recent years several attachment theorists and developmental psychologists have put forward different versions of the general hypothesis that – in the case of humans - there is an inherent causal and functional link between the quality of early infant attachment on the one hand, and the development of the ability for mentalizing on the other. In particular, the acquisition of mindreading skills by about four years (as indicated by passing the standard verbal false belief attribution tasks, Wimmer and Perner 1983; Wellman *et al.* 2001) has been suggested to be facilitated by the security of early attachment relationships (e.g., Hobson 2002; Fonagy and Target 1997; Fonagy *et al.* 1997ab; Meins *et al.* 1998, 2003). It has been proposed that
secure attachment and early parental interactive behaviours that correlate with infant attachment security at 12 months (such as maternal mind-mindedness at 6 months, Meins et al. 2001) play an important causal role in fostering the development of mentalizing as they facilitate the early appearance of explicit mindreading skills in preschoolers (e.g., Fonagy 2001; Fonagy and Target 1997; Fonagy et al. 1997ab; Meins et al. 1998, 2002, 2003). The supporting evidence consists mainly of correlational findings showing that secure infant attachment at one year (as well as its earlier correlates such as maternal mind-mindedness at six months) predict earlier ability to pass the standard verbal false believe tasks around four years (e.g., de Rosnay and Harris 2002; Fonagy and Target 1997; Fonagy et al. 1997ab; Meins et al. 1998, 2003; Ontai and Thompson 2002).

However, there are a number of empirical and theoretical reasons to question the strong form of the hypothesis that postulates either a specific evolutionary role and/or a direct ontogenetic causal developmental 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. First, one problem lies in the fact that the correlational findings linking infant attachment security during the first year to earlier success in passing the explicit theory-of-mind tasks have not been supplemented with specific causal developmental models. Ideally, such models should spell out the particular ways in which the secure quality of preverbal attachment and/or factors such as parental talk involving mental state attributions to the baby at six months exert their facilitating effects on the acquisition of the cognitive skills involved in inferring, attributing, representing and reasoning about epistemic mind states of others (such as beliefs). Numerous plausible causal linking factors have been suggested in this regard that, however, exert their effect later in development. These include the frequency of
perspective taking in caregiver-child linguistic exchanges (e.g., Harris 2005; de Rosnay and Hughes 2006; Lohmann et al. 2005), the amount of complementary role play in the family, the degree of family talk about conflicting emotions and differential conflict- and emotion-regulation strategies (Vinden 2001; Cutting and Dunn 1999; Sabbagh and Callanan 1998; Dunn et al. 1991; Ruffman et al. 1999), and the number of references to thoughts and beliefs and the relationship specificity of children’s real-life accounts of negative emotions (Hughes and Dunn 2002). The independent effects of these (and other) factors on the earlier appearance of explicit theory-of-mind performance around four years are well documented (for a review, see Fonagy et al. in press). Other facilitating factors that are unlikely to be related to infant attachment security include having older siblings (Perner et al. 1994; Ruffman et al. 1998), the relative frequency and precocity of using complex syntactic complement structures (Astington and Jenkins 1999; de Villiers and de Villiers 2000, de Villiers and Plyers 2002), or the maturational development of executive function, selective attention and response inhibition capacities (Leslie 2000; Leslie et al. 2004; Perner and Lang 2000).

Second, the demonstrated facilitating effects of early secure attachment on theory-of-mind development are not strong. Moreover, in several studies they only show up in theory-of-mind tasks that involve reasoning about both beliefs and emotions (such as belief-desire tasks) in contrast to tasks probing purely epistemic states such as false beliefs (e.g., Fonagy et al. 1997a; see Fonagy et al, in press, for a review). Furthermore, there are strong recent arguments and supporting evidence showing that the standard use of explicit verbal false belief tasks (Wellman et al. 2001) to diagnose the developmental achievement of ‘full’ mindreading skills around age four involve several complex linguistic, pragmatic interpretational, and executive function requirements. Thus, it may be that young children have difficulty passing false belief tasks not
because they still lack the capacity to mentalize (or at least some components of theory-of-mind), but rather due to deficits with some of the non-theory-of-mind related component skills that these tasks require (Bloom and German 2000; Leslie et al. 2004). This may have seriously distorted the validity of the generally accepted developmental time-table for the ontogenetic establishment of mentalizing abilities. For these reasons, some researchers have recently argued explicitly for the abandonment of the standard false belief tasks as a general diagnostic tool for establishing the attainment of theory-of-mind skills (e. g., Bloom and German 2000).

Complementary to these methodologically-based doubts, there is converging recent evidence (based on paradigms not plagued with the methodological problems outlined above) demonstrating mentalization much earlier in human ontogeny. For example, Tomasello and his colleagues have shown that infants exhibit a variety of theory-of-mind abilities as early as 12 months. These include spontaneously monitoring others’ perceptual access to transformations of relevant aspects of reality, inferring and attributing to them the corresponding epistemic mental states of informedness or ignorance, and relying on such mental attributions to modulate one’s own goal-directed behaviors in the presence of the other (e. g., Tomasello and Haberl 2003; Tomasello et al. 2005). Similarly, there is convergent new evidence based on implicit non-verbal versions of the standard false belief tasks using violation-of-expectation looking time paradigms showing that infants, as young as 13-months of age, monitor others’ perceptual access to reality, infer, attribute, and represent their true or false beliefs about the situation accordingly, and predict the other’s subsequent behavior by relying on the representational content of such attributed mental states (Csibra and Southgate 2006; Onishi and Baillargeon 2005; Onishi et al. 2007; Southgate et al. in press; Surian et al. in press).
Finally, recent evidence from comparative animal literature calls into doubt the notion that the basic ability for implicit mindreading is an exclusively human-specific capacity. These studies suggest that some rudimentary and possibly highly domain-restricted forms of mentalizing seem to have evolved independently and convergently in a number of non-human social species as well (such as apes, goats, and avian species including crows, ravens and scrub-jays, see Bugnyar and Heinrich, 2005; Emery and Clayton 2004; Hare et al. 2001; Kaminski et al. 2006; Tomasello et al. 2003).

To conclude: Based on this brief review, we argue that the currently popular view endorsed by a number of attachment theorist and infant researchers that assumes a - possibly evolutionary-based and human-specific - direct causal and functional link between early security of infant attachment on the one hand, and the acquisition of explicit mentalizing skills on the other, must be significantly revised. The relevant developmental and comparative evidence of recent years seems more compatible with the alternative view that an implicit and automatic capacity for mentalizing about 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 in humans. Furthermore, this capacity for implicit mentalizing is unlikely to be a uniquely human competence either as, in its basic form, it seems to have also evolved independently and convergently in a number of non-human social species as well that live in highly competitive niches.

The above does not imply that certain types of early interactive patterns characteristic of specific attachment relationships of human infants would have no developmental effect whatsoever on one’s later ability to functionally and efficiently use the
capacity for mentalizing in coping with interpersonal interactions and relationships. There are, in fact, good empirical reasons to believe that certain dysfunctional types of early attachment relations involving severe neglect, abuse, dissociative, highly intrusive, or grossly unpredictable patterns of parental reactivity have significant and long-term detrimental and disruptive effects on one’s later capacity to functionally use the innate competence for on-line mentalization as an adaptive interpersonal coping strategy (see e. g., Fonagy et al. 2000, 2002, 2003, in press; Gergely et al. 2002).

In this chapter we advance the argument that the specialized capacity for mentalization is likely to be an independent social-cognitive adaptation whose primary evolutionary function is separate and unrelated to the basic evolutionary function for which the attachment instinct system has been selected. If so, it would not be surprising if security of early infant attachment in humans did not turn out to play a direct causal functional role in the ontogenetic development of our mentalizing capacity. Nevertheless, we propose that when maladaptive patterns of affective parental reactivity become the dominant features of the re-occurring interactive structure of the infant’s primary attachment relationships, they can and do play a significant causal role in pathologically undermining the developing self’s potential to rely on its innate mentalizing capacity as its dominant social-cognitive strategy to cope with interpersonal situations and intimate and affiliative relationships during later life. Elsewhere we have argued (and analyzed a specific clinical example to illustrate, Gergely and Unoka in press) that the ensuing inhibitory or distorting effects on the self’s later capacity to mentalize stem from the continued influence that the self’s non-conscious and procedurally represented primary affective self-other relationship representations exert on perceiving, anticipating, and interpreting the behavioural and affective reactions of interactive partners. In the rest of the chapter, we elaborate an evolutionary-based
social developmental model about the structural nature and ontogenetic origins of such primary affective relationship representations. In particular, we shall specify the role of representation-building and attention socialization mechanisms that result in the early establishment of such affective self-other representations that capture the causal properties of the re-occurring patterns of specific attachment interactions during early development.

11.2 The social construction of the representational affective self in humans

11.2.1 The ‘constitutional self’: The evolutionary function of basic emotions, their modular organization and pre-wired causal properties

We assume that the infant’s innate ‘constitutional self’ (Gergely 2007a; Gergely and Watson 1966, 1999; Fonagy et al. 2002; in press) has a rich pre-wired affective structural organization. Apart from genetically-based individual temperamental differences (Kagan 1994; Rothbart 1989), we assume – along with current-day evolutionary theories of emotions (Ekman 1992; Ekman et al. 1972; Izard and Malatesta 1987; Tomkins 1995; see also Darwin 1872) - that the constitutional self also contains innate specifications of the core physiological, motor, and functional components that make up the organizational structure of basic emotion programs (such as fear, anger, joy/interest, disgust, and sadness). Current cross-cultural and experimental research has identified about six basic categorical emotions that are characterized by specific and differential patterns of physiological bodily arousal states on the one hand, and/or motor programs of emotion expressive facial displays patterns on the other. These are universally recognized and cross-culturally shared by adult humans around the world (Ekman 1992, Ekman et al. 1972). The basic emotions are types of adaptations that originate from similar automatic
emotional coping systems present in several non-human social animal species such as our primate ancestors (Pankshepp 1998; Suomi 1999).

Basic emotion programs are pre-wired stimulus-driven and procedurally represented behavioral automatisms (Ekman 1992). They have evolved to generate quick and adequate automatic coping responses (e.g., fight or flight) to deal with the specific types of environmental challenges that their causal triggering conditions represent (e.g., predator threat, territorial conflicts, mating competition, etc.). Such largely invariant, stereotypic and reoccurring environmental stimulus conditions became represented through evolution as innately specified causal inputs that automatically activate the basic emotions. These primary emotion programs consist of two major structural components: a) a specific pattern of physiological bodily arousal state and b) pre-wired emotion-specific motor routines. The latter component consists of two types of fixed behavioral automatisms: (i) stereotypic action tendencies (approach/avoidance, fight/flight, etc.) and (ii) expressive emotion-specific fixed facial-vocal display patterns and bodily postures. The evolutionary function of both components is to provide specific coping responses with the built-in goal to adaptively change (i.e., to eliminate, modify, or maintain) the environmental triggering conditions so that the challenges represented by these triggers are altered and thereby resolved.

In non-human animal species the activation of basic emotions is largely stimulus-driven and automatic, and under the external control of pre-specified inputs. Furthermore, the range of causal consequences that the execution of their fixed behavioral components (e.g., attack, threat displays, courting behavior, distress calls) can bring about in the animal’s environment is also highly restricted to a small set of stereotypic outcomes. This is so because the scope of their social environmental effects is restricted to the equally small number of innately pre-specified
stereotypic reactions they can induce in other conspecifics (e.g., flight/fight, submission, acceptance/refusal of mating approaches). Therefore, the ensuing stereotyped emotional interactions among conspecifics transform the original causal environmental emotion-triggering conditions in a relatively small number of largely invariant ways. The consequent environmental changes then induce either automatic termination of the organism’s basic emotional arousal, or inhibition of its behavioral expressions (due to modifying its original triggering conditions), or activation of a different basic emotion (due to establishing new environmental triggering conditions).

Clearly, this type of innately specified, automatic and stimulus-controlled modular organization of basic emotions can fulfill its primary function described above only in evolutionary niches that are characterized by stable, reoccurring, and largely invariant social environmental conditions. Indeed, the specialized and modularly organized behavioral coping systems provided by the basic emotion programs must have evolved under the selective pressure that such invariant and stable conditions of the species’ evolutionary environment represented.

11.2.2 The emergence of context-sensitive modifiability of the control mechanisms of basic emotion programs in humans

Such highly stable evolutionary conditions, however, have ceased to characterize the increasingly complex and changing cultural, social, and technological environment of early humans. In human social interactions the same expressions of a basic emotion can evoke largely variable responses from different individuals as a function of their personality traits, social status, current dispositions, informedness about relevant aspects of the situation and their cognitive evaluation of the significance of the emotion-inducing context. During ontogenetic
development, individuals can also acquire variably large and flexible repertoires of non-stereotypic emotional response alternatives whose context-sensitive selective activation come to be – at least partially and some of the time - under their cognitive voluntary control. This results in large variability and significant individual differences in the range and types of reactions that the automatic behavioral expressions of a conspecific’s basic emotional arousal can induce in others under different situational and social contexts.

We suggest that the evolutionary emergence of such significantly growing variability and increase in the types of relevant social environmental conditions must have co-evolved with correspondingly significant modifications in the inhibitory and response selective control mechanisms of basic emotion programs in humans to ensure their continued role as functional coping systems. For the basic emotion programs to remain adaptive in the face of such conditions of situational and individually variable social environmental reactivity, their control systems must have evolved to incorporate mechanisms a) to evaluate, anticipate, and represent the variable causal consequences that specific emotional reactions are likely to result in under different interpersonal and situational contexts, and b) to selectively inhibit or flexibly modify their automatic emotion-induced behaviors and expressive displays as a function of their anticipated situation-specific causal effects.

11.2.3 Early maternal emotion-regulation of the infant’s primary affective states within the human attachment system

In early phases of ontogenetic development, however, the affective behavior of young infants provides no indication of the availability of such flexible and context-sensitive emotion-control mechanisms. Infants seem largely unable to modulate or inhibit the automatic motor
expressions of their basic emotional arousal states. Their affective state regulation is largely under the external control of the attachment figure’s adequate reactions (Gergely 2007a; Gergely and Watson 1996, 1999; Sroufe 1996; Tronick 1989; Fonagy et al. 2002; in press). If the caregiver can sensitively monitor and correctly interpret the infant’s automatic affective state signals, she can react to them with adequate contingent emotional expressions or instrumental coping responses to regulate the baby’s emotional arousal. This can be achieved by at least two different mechanisms. First, through the pre-wired direct physiological state-modulating effects of certain maternal attachment reactions (e.g., through the innate soothing effect that gentle bodily contact or the specific intonation pattern of motherese induce in babies; see Fernard 1985, 1992; Hofer 1995; Cooper and Aslin 1990; Polan and Hofer 1999). Or second, through instrumental actions transforming the baby’s proximal environment, thereby appropriately changing the external conditions that have triggered (or would continue to maintain) the infant’s emotional arousal.

Through this process the infant’s basic emotion programs start to become modified to increasingly enable the young child to anticipate and cope with the characteristic situational variability of the habitual causal consequences that his or her specific emotional reactions typically lead to. This requires the establishment of experience-based representations of the expectable causal consequences that different emotion responses tend to induce from particular attachment figures under specific situational contexts. Note, however, that to enable the construction of such context-sensitive emotion representations during different person- and situation-specific affective interactions, the infant must possess a suitable event-monitoring and representation-building mechanism that can register, analyze, and represent the characteristic
causal dependencies between the self’s particular emotion responses and the specific reactions they invoke in others.

11.2.4 The causal contingency detection mechanism and its role in building context-specific primary emotion representations

Elsewhere we have proposed that human infants are, indeed, equipped with a specialized innate information processing and representation-building device, the so-called ‘contingency detection mechanism’ (Gergely and Watson 1996, 1999; Watson 1994) that, when applied to the domain of emotional interactions, can construct the required type of experience-based and context-sensitive primary emotion representations. This mechanism is specialized for the detection and representation of causal contingency relations between the infant’s specific responses and consequent environmental outcomes by registering the degree of causal dependency (in terms of conditional probability of co-occurrence over time) between responses and stimulus outcomes within particular situational contexts (see Gergely 2002, 2004, 2007a; Watson 1985, 1994, 1995, 2001; Gergely and Watson 1996, 1999; for technical details and review of supporting evidence). One basic developmental function of the contingency detection device is to identify and represent the degree of the infant’s causal control over different responses of the social environment and the specific contextual factors modulating these causal effects (Watson 1995, 2001). The resulting experience-based primary representations of the degree of causal efficacy of the infant’s responses form the basis of the young child’s developing primary sense of self-efficacy and social self-agency (Gergely 2002; 2007a; Gergely and Watson 1996).
Here we shall not describe the details of the structural and functional characteristics of this specialized causal representation-building mechanism (the interested reader is referred to Watson 1985, 1994, 1995; 2001; Gergely and Watson 1996, 1999; or Fonagy et al. 2002). For our present purposes we call attention to only one of its crucial properties: To discover and represent the actual degree of causal control that a specific infant behavior has over a particular response by the caregiver requires the monitoring, assessment, and representation of two different (and independent) aspects of the causal dependency relation that are obtained between these two target events over time. In other words, the contingency detection mechanism must construct two separate representations. The first must monitor the occurrences of a specific target behavior of the self and represent the degree of causal likelihood that it is followed by a particular monitored response of the other. The second must monitor the occurrences of the target response of the other and represent the degree of causal likelihood that it was preceded by the self’s target behavior monitored. Only the aggregate of these two values can provide a correct estimate and representation of the actual degree of causal relatedness of the two target events monitored (Watson 1994; Gergely and Watson 1996).

Let us clarify this point by an example (cf. Watson 1995). Imagine a frightened boy whose negative emotional arousal automatically activates a fear-expressive facial display towards her mother. Assume that she promptly responds to the infant’s emotion-expression by picking him up. By monitoring the causal structure of this specific type of attachment interaction over re-occurring episodes, the infant’s contingency detection mechanism will register that his fear-expressive facial-vocal display results in being picked up by mother, say, 80% of the time. On this basis alone, it would seem safe to conclude that the degree of causal efficacy of this specific attachment behavior in inducing the desired maternal response is, in fact, quite high. But
would this provide a correct estimate of the *actual* degree of the infant’s causal control over the mother’s specific response? Not necessarily. Imagine that this mother is over-controlling, or physically intrusive, or has an anxious, worried, and fearful disposition (projecting these emotions onto her baby), or is emotionally unstable and insecure with a constant need for physical proximity herself. Any of these maternal dispositions can result in highly frequent occurrences of picking her baby up even when this was *not* preceded (and evoked) by her infant’s fear-expression. To take into account such occasions when the maternal response was *not* brought about by the infant’s fear-expressive behavior, the contingency detection mechanism needs to (and does, see Watson 1985, 1994; Gergely and Watson 1996, 1999) separately monitor the occurrences of the maternal target behaviour, checking the conditional probability with which her reaction was actually preceded (and invoked) by the infant’s fear-expressive display. Assume that this second type of assessment indicates that only 20% of the mother’s target behaviors were, in fact, preceded by her infant’s fear expression. Clearly, under such circumstances it would be grossly misleading to represent the actual degree of causal control that the infant’s fear-expression has over being picked up by her mother as being very high.

In fact, as the evidence indicates (e.g., Watson 1985, 1994), the design structure of the contingency detection mechanism is sensitive to this problem and solves it by separately monitoring and representing the two types of target events and their respective degree of (prospective and retrospective) causal dependency in relation to each other. This results in setting up two separate representations for the two behavioral events: One that represents *the monitored target behavior of the infant and its degree of causal efficacy in bringing about the particular target response of the other* (this is the so-called ‘sufficiency index’); and one that represents *the monitored target response of the other and its degree of causal dependence on the previous
occurrence of the specific target behavior of the infant (this is the so-called ‘necessity index’, Watson 1994; Gergely and Watson 1996).

11.2.5 Two types of primary emotion representations of early affective attachment interactions and their causal properties: Self-to-Other versus Other-to-Self emotion schemes

Let us examine what kind of experience-based primary emotion representations are constructed when the contingency detection mechanism is applied to re-occurring affective infant-caregiver attachment interactions. There are three features of the resulting primary affective self-other representations that we think are of central importance because of the functional role they will play in supporting the infant’s emerging capacity for emotional self-regulation and control during affectively charged attachment interactions. First, they will represent crucial information about the characteristic causal properties of the specific emotion responses that are monitored (representing both their causal triggering conditions and their expectable consequences). Second, the characteristic profile of these causal conditions will be separately represented for different situational and person-contexts in which the specific target responses tend to occur. Third, separate primary emotion schemes will be established to represent the causal and situational properties of the self’s own emotion responses on the one hand, and those of the monitored target responses of different caregivers on the other.

Clearly, it can be highly adaptive for the infant to actively monitor and separately represent the causal and situational conditions under which different caregivers tend to produce specific types of responses. This is particularly relevant when such target behaviors of the other represent significant challenges for the self’s physical well-being and emotional state of arousal. Such representations provide vital information for the infant to anticipate emotionally highly
charged and potentially self-endangering interactions with particular caregivers in specific attachment situations and can therefore facilitate the generation of adaptive anticipatory coping reactions to deal with them.

Thus, we propose that the contingency detection mechanism creates two basic kinds of experience-based primary emotion representations that capture the causal properties of reoccurring types of affective interactions with particular attachment figures under specific situational contexts:

- **Self-to-Other causal emotion schemes**: These represent two types of predictable causal consequences that a specific emotion reaction is likely to lead to when performed in the presence of a particular attachment figure in a given situational context: (i) The relative likelihood that the infant’s emotion-expression will induce a particular response by the caregiver in the situation, and (ii) The probable consequences that the infant-induced reaction of the caregiver will give rise to in that situation: i.e., the ways in which it is likely to modify the environmental conditions that had triggered the emotional reaction in the first place.

- **Other-to-Self causal emotion schemes**: These represent the characteristic causal properties of specific target behaviors that are likely to be performed by the caregiver under certain attachment contexts. These behaviors of specific caregivers become active targets of monitoring either because they belong to the likely consequences of some of the infant’s own monitored emotion-expressive responses (by her Self-to-Other emotion schemes), or because their occurrence is likely to trigger a basic emotional reaction in the infant (thus, representing a significant environmental challenge for the self).

Such Other-to-Self emotion schemes will include causal information about the types of environmental triggering conditions that are likely to induce (and, therefore, can be used to
anticipate) the other’s monitored target behaviors. Importantly, such causal conditions may also include the infant’s representation of his or her own specific emotion-expressive responses as having triggered the caregiver’s monitored reactions in the past during particular situations and with differential likelihood. In such cases the infant can use this represented causal information to control, modify or pre-empt the occurrence of the other’s anticipated target reaction by inhibiting its potential cause; that is, by suppressing or modifying the behavioral expression of his or her own emotion reactions that are represented as likely to induce the other’s – potentially self-endangering - target behavior whose occurrence the infant is trying to avoid.

11.2.6 Primary response inhibition and substitution mechanisms of automatic emotional self-control

Two basic types of emotion control processes (involving primary automatic vs. secondary cognitive control mechanisms) serve affective self-regulation and coping in humans during emotionally charged interactions. We discuss primary automatic response inhibition first.

In our view, the basic developmental function of the infant’s experience-based primary emotion schemes is to provide the infant with representational means to anticipate and adaptively influence the course of affective attachment interactions with particular caregivers in the service of emotional self-control and coping. By extracting and representing the context- and person-specific causal patterns of affective interactions, these primary representations equip the infant with a repertoire of specific emotion schemes that can be used to anticipate and to some degree actively modulate the emotional consequences that particular types of attachment interactions will bring about.
There are two types of mechanism through which the infant’s primary emotion control system can utilize the representational contents of the Self-to-Other and Other-to-Self emotion schemes for these purposes. The first is a primary response inhibitory control mechanism that can automatically suppress the behavioral expression of a basic emotional arousal state. One condition that leads to the triggering of such a response-inhibitory process is when the activation of the self’s basic emotion reaction primes a specific Self-to-Other emotion scheme that represents the behavioral expression of the activated emotion state as likely to lead to severe negative consequences for the self in the given attachment context. In such cases, the anticipatory activation of the represented negative consequences exerts a direct and automatic inhibitory effect blocking the impending motor expression of the activated emotion thereby avoiding the realization of the negative consequences anticipated. This type of primary response-inhibitory emotion control mechanism is automatic and procedural in nature.

Evidence from attachment research (Cassidy and Shaver 1999) indicates that this primary response-inhibitory coping mechanism becomes operational rather early. An example is the separation-induced response pattern observed in the Strange Situation Test (Ainsworth et al. 1978) in a sub-group of one-year-olds who show avoidant attachment to their caregiver. Unlike other (securely or insecurely attached) infants, avoidantly attached babies typically do not exhibit the behavioral-expressive signs of distress or protest that would indicate their negative physiological arousal induced by separation from their caregiver (Ainsworth et al. 1978; Cassidy and Shaver 1999; Goldberg et al. 1995). In spite of this apparent lack of emotion-expressive reactivity to separation, physiological (heart-rate and cortisol-elevation) measures (e.g., Spangler and Grossmann 1993; Cassidy and Shaver 1999) indicate that – similarly to infants belonging to
other attachment categories - separation from the caregiver induces clear bodily indices of high negative emotional arousal and stress in such infants.

Importantly, this automatic inhibition of the expression of separation-induced distress is relationship-specific: no strong correlation has been found between the types of attachment infants exhibit towards different caregivers (see Cassidy and Shaver 1999; Fonagy 2001, for reviews). There is evidence (Ainsworth et al. 1978; Cassidy and Shaver 1999) that the establishment of the primary response-inhibitory coping strategy in avoidant infants towards a particular caregiver is related to (and is best explained as an adaptation to cope with) the predictable and systematically negative emotional reactivity habitually exhibited by the caregiver in response to the infant’s expressions of negative affect across situations (Watson 2001).

Note that the same automatic response-inhibitory control mechanism can also be triggered by the activation of some of the infant’s Other-to-Self emotion schemes, which represent the causal properties of those target behaviors of a given caregiver that are likely to induce negative emotional consequences for the infant in the given attachment context. This can happen when the situation triggers in the baby a basic emotional arousal whose behavioral expression is represented in the infant’s Other-to-Self scheme as one of the potential causes that are likely to invoke the self-endangering target behavior of the caregiver. In such cases the infant’s emotional arousal activates the Other-to-Self scheme in question leading to the anticipatory evocation of its representational content: the potential occurrence of the caregiver’s target behavior and its expectable negative consequences. This will trigger the automatic response-inhibition control mechanism to suppress the expression of the infant’s basic emotion arousal and thereby avoid the realization of its anticipated negative consequences by blocking their causal triggering event from occurring.
It should be pointed out, however, that, when exercised in isolation, this kind of automatic response-inhibitory coping reaction is of rather limited adaptive value for the infant’s emotion regulation for two reasons. First, while it automatically suppresses the expression of the activated emotion, it does so without affecting the physiological arousal reaction itself. One of the undesirable consequences is that while the inhibition of the emotion expression removes the source of likely further escalation of the infant’s emotional reaction, the negative physiological arousal already activated remains unmodified and unmodulated. Second, the inhibition of the emotion expression also leaves unchanged the environmental triggering conditions that led to the activation of the basic emotion reaction in the first place. Therefore, the infant’s emotional arousal will remain maintained as long as its environmental triggering source continues to be present. In short, such an automatic inhibitory “freezing” of the emotion response system can leave the self helplessly overwhelmed by its uncontrolled and continued state of heightened physiological arousal and stress that, if occurring frequently, can lead to deleterious long-term toxic consequences thereby resulting in the development of a rigid and dysfunctional organization physiological stress-regulation system (Francis and Meaney 1999; Francis et al. 1999; Fonagy et al., 2002; Pruessner et al. 2004; Wismer Fries et al. 2005).

The danger of toxic consequences could be coped with only by applying the second type primary emotion control mechanism that involves response substitution. This mechanism exploits the possibility provided by the representational contents of primary Self-to-Other and Other-to-Self emotion schemes to automatically activate a suitable substitute behavior that is represented as having a high likelihood to change the environmental conditions that had activated the infant’s original basic emotional arousal in the first place. Clearly, this can be
accomplished only if the infant possesses a rich and differentiated enough repertoire of response alternatives suitable to serve as adequate substitute responses in the given situation.

11.2.7 Secondary cognitive appraisal mechanisms and top-down voluntary processes of emotional-self control

We have argued that the automatic primary emotion control mechanisms – when applied over a sufficiently differentiated repertoire of primary emotion representations - allow the infant to exert a certain measure of adaptive behavioral control in the service of emotional self-regulation during early affective interactions. However, this primary emotional coping system is severely limited both in its scope and its flexibility. This is so because the Self-to-Other or Other-to-Self emotion schemes are stimulus-driven and procedurally represented emotion representations that become automatically activated whenever the stimulus situation is sufficiently similar to their represented environmental triggering conditions or their associated situational and person-contexts. This, by necessity, restricts their scope of adaptive applicability to the narrow domain of the reoccurring types of affective attachment interactions with caregivers, whose features are represented by the primary emotion schemes in the first place.

However, as children develop, they encounter an increasing variety of new interpersonal contexts of affective interactions often involving unfamiliar persons and situations. This results in two basic problems of adaptation that the primary emotion representations and their automatic emotion control mechanisms cannot satisfactorily cope with and that necessitate the establishment of an additional type of secondary top-down cognitive mechanism of emotion control (e. g., Lazarus 1991, Lazarus and Lazarus 1994; LeDoux 2000; Thompson 1994; Ellsworth and Scherer 2003; Ochsner and Gross 2005; Posner and Rothbart 2000).
The first problem arises when the new interactive situations encountered represent significant challenges for the self that would call for adaptive emotional coping responses, but that – due to their unfamiliarity and lack of sufficient similarity to the triggering conditions represented in the child’s primary emotion schemes - do not automatically trigger a basic emotion reaction. To recognize the realistic nature of the challenge that such situations represent for the self (and, consequently, to generate appropriate emotional coping reactions to them), the child has to rely on *secondary cognitive appraisal mechanisms*. By having access to an increasingly large pool of relevant knowledge and being able to derive relevant inferences from situational cues these top-down cognitive appraisal processes can generate realistic evaluations of the significance that these new types of input conditions represent for the self. The interpretation that such secondary cognitive appraisals provide can then be used by top-down executive action control systems to activate relevant emotion responses to cope with the situational challenge thus recognized.

The second type of problem arises when the surface features of the current interactive context automatically trigger the self’s primary and procedural emotion representations leading to non-adaptive and inappropriate emotion reactions. This can occur whenever the environmental situation shows sufficient similarity with the triggering conditions or associated situational and person-contexts that are represented by some of the self’s primary Self-to-Other or Other-to-Self emotion schemes. Since the input specifications of such primary emotion schemes represent the reoccurring patterns specific to the self’s early attachment interactions, their ‘spurious’ triggering by a featurally partially overlapping current context is highly unlikely to generate an automatic emotional reaction that is adaptive in the present situation. Therefore, if left cognitively unchecked, such automatic and similarity-driven contextual activation of the procedurally
organized primary emotion schemes are likely to result in inappropriate and reality-distorting projective constructions of the emotional significance of the current interpersonal situation. In order to control such dysfunctional automatic activation of primary emotion schemes, the self must again rely on top-down processes of secondary cognitive re-appraisal of current situational cues (e.g., Lazarus 1991; Thompson 1994; Ellsworth and Scherer 2003; Ochsner and Gross 2005) to inhibit and override the impending execution of the spuriously activated primary emotions. This ontogenetically developing capacity for top-down control over one’s primary emotional automatisms relies on two related aspects of cognitive processes of secondary evaluation. Both result in a re-interpretation of events and the generation of alternative response options.

First, the secondary re-appraisal processes can re-interpret the actual significance for the self of the causal evoking conditions that have automatically activated the primary emotional reaction in the first place. Based on stored or situationally inferred relevant information, one can re-appraise and revise the interpretation of what the emotion-triggering events signify for the self. This top-down cognitive re-evaluation can override the automatic output of the ‘quick-and-dirty’ stimulus-driven primary perceptual appraisal processes (LeDoux 1995, 1996, 2000; Ochsner and Gross 2005) that led to the original automatic triggering of one’s primary emotion reaction. Second, the secondary cognitive evaluative processes can also re-appraise the realistic significance of the expectable causal consequences that the automatic expression of the spuriously activated primary emotional reaction would lead to.

There are several developmental, representational, and cognitive preconditions, however, that need to be established before such secondary cognitive re-appraisal processes can be successfully applied to adaptively serve their emotion-regulative function. First, top-down
cognitive processes of emotional self-control need to be ‘alerted’ when a specific emotional arousal state has been activated. Therefore, the cognitive emotion control system must be able to monitor, detect and introspectively access internal state cues that signal the onset of a basic emotional reaction in the self. Second, this introspective emotion recognition must be based on sufficiently differentiated internal state signals to allow for the correct identification and self- attribution of the particular emotion category that has been triggered. Third, the cognitive processes of secondary re-appraisal must have access to sufficient and relevant (stored or inferred) information in the light of which adaptive re-evaluation of the realistic significance of the causal conditions of the basic emotion reaction can be successfully performed. Importantly, these processes of representational access and situational re-appraisal must be implemented under time pressure to deliver their output before the execution of the stereotypic behavioral components of the activated basic emotion program would be automatically implemented.

11.3 The establishment of second-order emotion representations:
Developmental mechanisms and social environmental preconditions

11.3.1 The development of the subjective sense of a differentiated affective self

Many currently popular developmental and attachment theories consider the human- specific characteristics of early caregiver-infant attachment interactions (see section 11.1 above) as evidence supporting a ‘rich’ mentalistic interpretation of such interactions in terms of primary ‘intersubjectivity’. These theories assume that already during the first months of life human infants have introspective access to a variety of differential mental states (such as emotions, goals, intentions, or motives), can recognize and attribute such internal states to the other’s mind, and experience subjective and mutual awareness of ‘being in’ and ‘sharing’ such internal states
with the caregiver (see Trevarthen 1979, 1993; Trevarthen and Aitken 2001; Braten 1988, 1992, 1998; Meltzoff 2002; Meltzoff and Gopnik 1993; Meltzoff and Moore 1977, 1989, 1998; Hobson 1993, 2002; Stern 1985; Giannino and Tronick 1988). Elsewhere we have criticized this primary intersubjectivist interpretation on a number of grounds arguing that it attributes overly ‘rich’ mentalizing capacities to very young infants that are not necessary to explain the characteristics of early mother-infant interactions in humans (see Gergely 2002, 2007a; Gergely and Watson 1996, 1999; Fonagy et al. 2002 in press, for details). We proposed an alternative model of early socio-emotional self development (Gergely and Watson 1996, 1999; Fonagy et al. 2002) within the framework of contingency detection and attachment theory (Watson 1985, 2001). We hypothesized that infants’ subjective awareness of differentiated emotional self states has important social interactional origins with an individually variable developmental course of ontogenetic unfolding. In this view, subjective self-awareness and sense of differential affective self states is viewed as a (relatively hard-won) developmental achievement rather than being a universally given pre-wired capacity. Similarly, we see the establishment and ‘intersubjective’ recognition of mutual ‘sharing’ of affective states with others as an emergent property of the subjective self, rather than an initial and universal starting state of human infants’ mental life.

In particular, we have proposed that the subjective sense of discrete emotional self states (e.g., awareness of being ‘angry’, rather than just experiencing an undifferentiated negative state of tension) is established as a result of infants’ repeated experience with the pattern of contingent parental reactivity and social „mirroring” feedback reactions that the automatic expressions of these - initially non-conscious – basic emotional arousal states evoke from the attachment environment. Our model is based on two central assumptions:
1. The introspectively inaccessible constitutional affective self. We assume (together with numerous others) that the infant’s innate constitutional self has a richly structured internal organization at birth. While on the one hand the constitutional self is characterized by significant genetically-based individual temperamental differences, it also contains a basic set of pre-wired universal categorical emotion programs that are primary biological adaptations (Ekman 1992; Ekman et al. 1972; Gergely and Watson 1996). As argued earlier, these basic emotions are best understood as pre-wired, stimulus driven, procedural physiological and motor automatisms that are initially not accessible to conscious awareness and over which the baby has no voluntary control at first. Early affect-regulation is carried out mainly by the attachment environment, as the caregiver, reading the infant's automatic emotion expressions, reacts to them with appropriate affect-modulating interactions and emotion displays (Gergely 2007a; Gergely and Watson 1996, 1999; Fonagy et al. 2002). Thus, while infants may be sensitive to the general (positive vs. negative) hedonic quality of their affective states, we assume that they have no initial introspective awareness of their discrete basic emotion states as categorically distinct subjective self states (Gergely and Watson 1996). It is in this sense that we consider the constitutional affective self to be initially introspectively inaccessible: while it contains discrete and categorical basic emotional states, these pre-wired primary automatisms are at first not subject to introspective awareness and so they do not present themselves as subjectively experienced differential internal self states.

We further hypothesize (contrary to classical approaches, such as Bruner et al. 1966; Mahler et al. 1975; cf. Gergely 2000) that in human infants due to a pre-wired directionality bias the attention system is initially dominantly externally oriented showing a primary sensitivity-bias towards exteroceptive stimulation. Therefore, it is assumed that very little introspectively
directed attentional monitoring of internal self states takes place at the earliest months of life (see Gergely and Watson 1996, 1999; Fonagy et al. 2002, for supporting arguments).

2. The social interactive origins of affective subjectivity: the development of introspectively accessible emotional self-representations. We assume that for the categorical basic emotion states of the constitutional self to become introspectively accessible and subjectively differentially experienced, two basic conditions must be established developmentally. On the one hand, the primary and procedural basic emotion programs need to become associated with second-order representations that, when activated, are cognitively accessible to introspectively oriented attentional self-monitoring processes. In other words, we suggest that the initially non-conscious primary emotion states become introspectively detectable differentiated states of subjective awareness through the activation and cognitive access of their associated second-order representations only. For this to happen, it is also necessary that the primary orientation bias of the young infant’s attention system towards external stimulus events should become modified. The infant’s attentional system must become partially directed towards the self’s internal states to allow the active introspective monitoring and detection of the activation of the self’s second-order emotion representations. Introspective emotion recognition then becomes possible whenever the second-order emotion representations are triggered by the stimulus-induced automatic arousal of the self’s basic emotions to which they are associatively linked. This leads to the introspective self-perception of ‘being in’ a particular differentiated emotion state giving rise to the discrete subjective awareness of that state.

The cognitively accessible and introspectively monitored second-order representations of the primary emotion states can then provide the basis for the establishment of a - continuously
updated - representation of the dynamically changing affective states of the self. This makes it possible to generate self-predictions about what kind of emotion-induced behavior one is about to perform by anticipating the impending execution of the emotion-specific action schemes that the self’s activated basic emotional arousal state is about to trigger. To become subjectively aware of one’s current dispositional affective states before the automatic activation of the primary emotion-induced action-tendencies would take effect is a central precondition for the self’s emerging capacity to exercise on-line affective self-control (Gergely and Watson 1996). The introspective detection of one’s emotional arousal state allows the self to foresee and cognitively evaluate the likely consequences that the anticipated primary emotion reactions would lead to in the given interpersonal context. The self can then avoid such non-adaptive outcomes by inhibiting or modifying the anticipated primary emotional reactions in time (i.e., before they would become executed) rather than being automatically driven to act them out.

Our developmental proposal is that both of these preconditions for the establishment of the introspectively accessible subjective self becomes fulfilled as a result of, and to the degree to which, the attachment environment provides systematic contingent feedback reactions to the automatic expressions of the initially non-conscious primary emotional arousal states of the infant. In other words, it is the repeated experience that one’s current internal states are externally “mirrored” or “reflected” through the infant-attuned contingent reactions of the social environment that leads to the development of a subjective sense and awareness of one’s differentiated affective self states.

Of course, no amount of social “mirroring” could result in the establishment of these two (representational and attentional) preconditions unless infants were equipped with an adequate mechanism to detect and represent the contingent relatedness between their automatic state-
expressive displays on the one hand, and the ensuing environmental mirroring reactions on the other. This mechanism must be able to internalize the representation of the external “mirroring” feedback signal as a second-order representation of the primary self state whose automatic expression invokes the contingent “mirroring” reaction in question.

In previous work on the social biofeedback theory of parental affect-mirroring (Gergely and Watson 1996, 1999) we have proposed that the infant’s contingency detection device has the necessary structural properties to establish second-order emotion representations when applied to contingent affect-mirroring interactions. As already discussed, this pre-wired causal contingency analyzer (Watson 1985, 1994) automatically monitors and assesses the degree of contingent relatedness between the infant’s responses and subsequent external stimulus events. By monitoring three different aspects of contingent relatedness in parallel (temporal contingency, spatial similarity, and relative intensity, see Gergely and Watson 1999, for a review) the mechanism can identify those aspects of the social environment that are under (some specific degree of) causal control of the infant’s state expressive displays. One important developmental consequence of this causal contingency analysis is that the resulting discovery of the degree of contingent control that the self exercises over aspects of the social environment generates a subjective sense of causal efficacy and self-agency in the infant.

11.3.2 The role of “markedness” of affect-mirroring expressions in the internalization of second-order representations of affective self states

One of the human-specific characteristics of infant-caregiver interactions (see section 11.1) is that infant-attuned caregivers are inclined to produce ‘marked’ affect-mirroring expressions that are induced by their baby’s automatic emotion displays. Such ‘marked’ affect-
reflecting expressions are *saliently transformed versions* of the stereotypic motor displays of categorical emotions that normally express the caregiver’s actual affective states (Gergely and Watson 1996, 1999; Fonagy et al. 2002). ‘Markedness’ involves motor transformations that result in exaggerated, slowed down, schematic, abbreviated or only partial execution of the normative display pattern of basic emotion expressions. Furthermore, marked affect-mirroring is typically accompanied by *ostensive-communicative cues* (such as direct eye-contact, eye-brow raising, vocal intonation pattern characteristic of motherese) that are also among the human-specific features of early caregiver-infant interactions (Csibra and Gergely 2006).

Our original functional interpretation of ‘markedness’ (Gergely and Watson 1996) emphasized its importance as a cue to signal that the displayed emotion “is not for real” and should therefore be “decoupled” from the mother: i.e., it’s attribution as her actual and real emotion state should be inhibited. Here we further propose that ‘marked’ affect-mirroring interactions constitute a specific form of ‘pedagogical’ manifestations of cultural knowledge transfer (Csibra and Gergely 2006) in the domain of emotion socialization.- This new theoretical perspective allows us to extend the original functional account of ‘marked’ affect-mirroring interactions in two ways: 1. We shall show why ‘marked’ affect-mirroring displays become *interpreted self-referentially by the infant* leading to their referential ‘anchoring’ as second-order emotion representations to the constitutional self’s primary emotions. 2) We shall explicate how the ‘markedness’ of contingent affect-mirroring expressions contributes to the modification of the primary directional bias of the infant’s – initially dominantly externally oriented - attention system resulting in its eventual *internally directed deployment to introspectively monitor the self’s internal affective states*. 
11.3.3 Natural ‘pedagogy’: The functions of cues of ‘ostensive communication’ and ‘referential knowledge manifestation’ in communicative knowledge transfer

Recently, Csibra and Gergely (2006) developed a new evolutionary-based theory of cultural transmission postulating a human-specific learning system that they call natural ‘pedagogy’ (Gergely 2007b; Gergely et al. 2007). They propose that humans have become adapted to spontaneously manifest relevant cultural knowledge to conspecifics who are naturally inclined to attend to and learn from such knowledge manifestations through a specialized cue-driven social communicative learning mechanism of mutual design. Human pedagogy ensures efficient intergenerational transfer of relevant cultural knowledge in a variety of knowledge domains including words, gestural symbols, artifact functions and novel means actions, valence properties of object kinds, social habits, rituals, etc. (Gergely 2007b; Gergely and Csibra 2005, 2006; Gergely et al. 2002, 2007; Csibra and Gergely 2006).

Here we argue that natural pedagogy is also employed in the domain of emotion socialization to identify, sensitize to, and transfer relevant knowledge about those categorical emotions to the infant that are culturally universal and shared among humans. In this view, early infant-caregiver affective interactions (involving ostensively cued ‘marked’ forms of contingent emotion-mirroring) constitute a special case of pedagogical knowledge transfer whereby sensitive caregivers establish second-order representations in infants that identify and encode the categorical emotions that are culturally universal and can be shared with and communicated to other humans.

The built-in mutual design structure of natural pedagogy involves biological preparedness both for providing and receiving relevant cultural information among humans. Knowledgeable caregivers show a natural inclination to express their ‘communicative intention’
(cf. Sperber and Wilson 1986) to transfer new and relevant knowledge to ignorant juveniles by addressing them through specific ‘ostensive’ communicative cues (such as eye-contact, eye-brow flashing, contingent reactivity, or the prosodic pattern of ‘motherese’) to which infants show pre-wired sensitivity and preferential orientation (see Csibra and Gergely 2006, for a review of the supporting evidence). Ostensive cues are followed by referential cues (such as gaze-shift) to help the infant identify the referent object about which new and relevant information is going to be demonstrated through the caregiver’s ‘marked’ forms of knowledge manifestations. These ostensive and manifestative pedagogical cues trigger in infants a specially receptive learning attitude that induces fast mapping of the manifested information to their representation of the referent.

Cues of ‘referential knowledge manifestations’ help the infant to identify and extract the content of the adult’s ‘referential intention’ (cf. Sperber and Wilson 1986), i.e., the new and relevant knowledge manifested about the referent, that the infant then encodes as a property of the referent kind. Such knowledge manifestations are cued by being performed in a specially ‘marked’ manner that involves a modified and saliently transformed version of the normative and primary form of application of the novel cultural knowledge skill that is being demonstrated for the infant to learn. (Imagine, for example, the difference between the ‘marked’ manifestative form of demonstrating to a novice ‘how to’ hammer a nail in versus the standard motor execution of the same routine when it is performed simply to fulfill its primary function of driving a nail in.) The features of ‘markedness’ that characterize knowledge manifestations serve several pedagogical functions. They cue the infant that the demonstrated content conveys relevant and culturally shared information about some essential property of the referent to be identified and represented. They also help the infant infer what aspect of the manifested knowledge should be
encoded by foregrounding and selectively emphasizing the new and relevant information that should be extracted and internalized as part of the referent representation (see Gergely 2007b, for details).

11.3.4 The role of cues of ‘ostension’ and ‘markedness’ in inducing the self-referential interpretation of ‘marked’ emotion-mirroring displays

We propose that the cues of ‘markedness’ of emotion-mirroring parental expressions constitute a special instance of the ‘marked’ forms of manifestative knowledge displays that characterize communicative acts of pedagogical knowledge transfer. ‘Marked’ affect-mirroring displays are also accompanied by ostensive cues that signal the adult’s communicative intent during pedagogical interactions (such as direct eye-contact, raised eye-brows, the vocal pattern of motherese, and addressing the infant by name). Above we argued (Gergely and Watson 1996, 1999; Fonagy et al. 2002) that as a consequence of these cues the infant inhibits the attribution of the expressed emotion to the caregiver as her ‘real’ feeling by referentially ‘decoupling’ the emotion display from her. We can now add that the cues of ostension and ‘markedness’ of emotion manifestation during affect-mirroring interactions also function to trigger the referential interpretive and receptive learning attitude of the infant’s ‘pedagogical stance’. In other words, these pedagogical cues induce the infant to search for and identify the intended referent of the caregiver’s ‘marked’ emotion manifestation and to ‘anchor’ the mirroring display to some referent other than the caregiver’s actual emotion state (from which it has been ‘decoupled’). In trying to establish the referent of the ‘marked’ emotion display, the infant will rely on cues of referent identification (such as eye-gaze direction) that accompany the caregiver’s affect-mirroring expression. Since the caregiver is looking at the infant while displaying her emotion-
reflecting expressions, the infant’s attention will be directed towards his or her *own* face and body as the likely spatial locus of the referent to which the ‘marked’ (and ‘decoupled’) emotion-mirroring display should be *referentially* ‘anchored’. Note that the infant’s automatic facial-vocal emotion expressions that are being mirrored by the caregiver emanate from the very same locus as well. The infant’s contingency detection mechanism identifies these emotion expressive responses of the self (and their proprioceptive bodily correlates) as the source of contingent control that the infant exercises over the caregiver’s mirroring response (see Gergely and Watson 1996). These two sources of information converge to identify the *infant’s own internal emotional arousal state and its motor expressions as the referent of the caregiver’s ‘marked’ mirroring display*. As a result, the infant will referentially ‘anchor’ the representation of the ‘marked’ mirroring display to his or her own primary and procedural emotion program and will internalize it as its second-order representation. In sum: This is how - through activating the infant’s pedagogical stance - repeated experience with ostensively cued ‘marked’ affect-mirroring feedback reactions of infant-attuned caregivers can ‘teach’ the infant about the self’s differential categorical emotion states that become represented as universal and cognitively accessible emotion categories shared by other humans. This is achieved by the referential ‘anchoring’ and internalization of the caregiver’s ‘marked’ mirroring displays as second-order emotion representations that become associated with the infant’s primary and procedural emotional self states. Through this process the infant’s attentional system becomes introspectively sensitized and internally directed to detect and monitor the activation of the self’s affective states. This (together with the sensitization effects of the contingency-based social biofeedback mechanism as described by Gergely and Watson 1996, pp. 1190-6) also results in heightened sensitivity to the internal proprioceptive cues that accompany emotional self-expressions. A further
consequence is the development of subjective awareness and emergent voluntary induction of the instrumental effects of contingent control that the self exerts through its emotion responses over the infant-attuned reactivity of the social environment.

Through these processes the infant’s social mirroring-environment contributes significantly to the establishment of the introspectively detectable subjective self by populating it with cognitively accessible second-order representations of internal affective self states. This provides the representational basis that, together with the infant’s introspectively socialized attentional system, make the developmental emergence of top-down cognitive emotional self-monitoring and secondary re-appraisal processes possible.

11.4 Mentalizing and its role in secondary cognitive re-appraisal processes of emotional self-control

Let us now turn to the important role that the ability to mentalize can play in enhancing the self’s developing competence to achieve top-down cognitive emotional self-regulation and control by increasing the efficiency and scope of the secondary cognitive re-appraisal processes involved. We speculate that the proper evolutionary domain (cf. Sperber and Hirschfeld 2004) of the cognitive adaptation for mentalising may have originally been restricted to infer and represent the causal intentional mental states of other minds only. This was sufficient to confer significant evolutionary benefits on the mindreader by allowing him or her to anticipate the likely actions of other conspecifics in primarily competitive situations. However, we hypothesize that with the human-specific establishment during early social and emotional development of cognitively accessible second-order representations of internal self states, the proper domain of the mentalizing capacity becomes ontogenetically extended to include in its actual domain the mind
of one’s own self as well. This is likely to result in significantly better coping strategies in the realm of competitive and cooperative interactions and affective relationships by enabling the self to anticipate its own actions and reactions as well as those of others, through introspectively detecting, self-attributing, representing, and reflecting about one’s own internal mental states that are the causal sources underlying one’s actions.

So how is the inferential system for mentalising applied introspectively in the domain of emotional self-regulation and control to make coping with affectively charged interpersonal interactions more effective? Clearly, one’s mindreading abilities can often provide highly useful and relevant information to instruct the self’s secondary cognitive emotion re-appraisal mechanisms. This would be the case, for example, when these top-down cognitive processes are applied to re-evaluate the realistic significance for the self of the causal input conditions that have induced one’s primary and automatic emotional arousal reactions in situations where these triggering conditions involve another person’s actions, reactions, or emotion expressions directed towards the self. In such cases mentalizing processes can support secondary cognitive re-appraisals by accessing and reasoning about stored knowledge concerning the other person’s enduring mental dispositions, self-related attitudes, long-term aspirations, temperamental and personality traits, or recent life events that are relevant for the reconstruction of the mental reasons behind the other’s self-directed actions. Similarly, one’s mentalizing ability can be recruited to detect and interpret current situational or behavioral cues that are relevant to infer what particular present desires, intentions, and (possibly false) beliefs about the situation and/or one’s own self may have induced the other’s responses towards the self that led to the triggering of one’s automatic basic emotional arousal response.
Mentalizing can also significantly facilitate the generation of secondary cognitive evaluations of the likely causal consequences that automatically acting out or expressively displaying one’s basic emotional arousal reactions would lead to in the given interpersonal context. The validity and scope of the predictions generated by such secondary cognitive evaluations can be qualitatively enriched by mentalistic reasoning about what kind of causal mental states and reactions would be expectably induced in the other person by the self’s automatic emotion-induced responses. Such mentalizing inferences can take into consideration a range of accessible or inferable information about relevant situation- and person-specific motivational factors such as the other’s temperamental features, personality traits, childhood history, current dispositional states, his or her quality of relationship with and available relevant knowledge about one’s own self, etc. Importantly, in generating predictions about the other’s likely reactions to one’s potential emotional outburst a good mentaliser may also be able to take into account available knowledge about the quality of the other’s mentalising skills that the latter can be counted on to apply when evaluating the reasons behind one’s own emotional reactions towards him or her in the given situation.

11.5 Conclusions

One of the arguably species-unique abilities of humans is to exercise on-line introspective control over their primary emotional impulses in affectively charged interpersonal situations and relationships. Individuals, however, differ significantly in the degree to which this remarkable capacity is available to them or can be put to functional use. In this paper we have outlined a social-cognitive developmental theory specifying the early social environmental preconditions and cognitive representation-building and attention socialization mechanisms that underlie the
ontogenetic unfolding of the introspectively accessible representational affective self. This approach relates the individual variability in the capacity for introspective access to affective self states and emotional self-control to the differences in the availability of a contingently reactive infant-attuned attachment environment in which the infant’s automatic emotion expressive responses induce contingent external feedback reactions and ‘marked’ forms of affect-mirroring displays by caregivers. We have argued that these human-specific features of early caregiver-infant interactions constitute necessary environmental input conditions for the infant’s representation-building and attention socialization mechanisms to establish second-order cognitively accessible representations for the initially non-conscious, stimulus-controlled and procedural basic emotional automatisms of the innate constitutional self. Our approach sheds new light on the functional role of the human-specific features of early affective attachment interactions (such as contingent turn-taking reactivity, ‘marked’ forms of affect-mirroring, or ostensive-communicative cues of ‘pedagogical’ knowledge manifestations) in early social-cognitive development. It interprets ostensively cued ‘marked’ affect-reflective interactions as a special case of ‘pedagogical’ knowledge manifestations (Csibra and Gergely 2006) that function to transfer relevant cultural knowledge about universally shared emotion categories through establishing second-order representations for the infant’s primary emotion states. We have contrasted this view with currently popular primary ‘intersubjectivist’ interpretations that consider the species-unique features of caregiver-infant interactions as characteristics of early secure attachment relationships that function as developmental precursors and facilitating conditions fostering the acquisition of mentalizing by early childhood. Finally, we have argued that the developmentally established second-order emotion representations allow the self to extend the original domain of the innate capacity for mentalizing about other minds to include
introspective mentalizing about one’s own internal affective mental states and anticipated emotion-induced actions as well. As a result the introspectively accessible representational affective self can employ its innate mentalizing ability in the service of emotional self-regulation and control by anticipating and adaptively modifying its own emotion-induced action tendencies to cope more efficiently with affectively charged interpersonal interactions.
References


Bigelow AE and De Coste C (2003). Infants’ sensitivity to contingency in social interactions with familiar and unfamiliar partners. *Infancy*, 4, 111-140.


Camras LA (1992). Expressive development and basic emotions. *Cognition and
Emotion, 6, 269-283.


Gergely G (2004). The role of contingency detection in early affect-regulative interactions and in the development of different types of infant attachment.


Polan HJ and MA Hofer (1999). Psychobiological origins of attachment and


