Foundations of human self-reflection: Error-monitoring developmentally predicts the emergence of self-representation

Cécile Gal[1], Dimitrios Askitis[1], Katarina Begus[1]
[1]Centre for Early Childhood Cognition, Faculty of Social Cognition, Department of Psychology, University of Copenhagen, Copenhagen, Denmark

Abstract
Self-reflection is central to human cognition and is arguably a defining feature of human consciousness. Yet, the mechanisms enabling our mind to reflect upon itself and to construct a conceptual sense of self remain poorly understood. This study brings empirical traction to the puzzle of self-reflection by studying human infants, across the period when a conceptual self is believed to emerge, from 12 to 18 months of age. Using a gaze-contingent match-to-sample paradigm with concurrent eye-tracking and EEG recordings, we find evidence of self-reflective capacities in 12-month-old infants: they exhibit a signature neural response of error-tracking (error-related negativity, ERN) after making incorrect as opposed to correct choices, and before receiving feedback. Remarkably, the magnitude of the ERN at 12 months developmentally predicted which infants will evidence a conceptual self-representation (mirror-self-recognition) 6 months later. This relationship was specific to internal error-monitoring (ERN), and not external feedback processing (FRN), and could not be explained by infants' age or general cognitive development. Furthermore, we show that only the group of infants, who later evidenced self-representation, exhibited an ERN, while FRN was evidenced in both groups. Moreover, the same group of infants who evidenced self-representation and internal error monitoring also adapted their behaviour post-error, by slowing their decisions and increasing visual exploration following mistakes. Thus, error-detection, if available, can be utilised for adaptive information-seeking and decision-making as early as 12 months of age. The finding that internal error-monitoring not only precedes but developmentally predicts the emergence of self-representation strongly supports the proposal that infants' error-monitoring and behavioural adjustments index rudimentary forms of metacognition. Moreover, these results point to the possibility that internal error-monitoring is functionally involved in building a conceptual self, thereby raising questions regarding possible factors modulating individuals' sensitivity to own errors and consequently their conceptual representation of themselves as distinct epistemic agents.

Adult human cognition is characterised by remarkable capacities for self-reflection. Our ability to take oneself as the object of thought, to explicitly and consciously reflect on what we know and feel, reminisce on our past selves, or imagine ourselves in the future has been of keen interest to philosophers and scientists for centuries, and is arguably a defining feature of human consciousness (*1*). Despite its centrality to human cognition, the mechanisms enabling our mind to reflect upon itself remain subject of much theoretical debate (*2*). One approach to tackling the enigma of human self-reflection, which has gained renewed interest in recent years (*3*), is to probe its form and processes by investigating its development. Conceptual sense of self is believed to be absent at birth and to develop from its implicit forms over the course of the first two years of life (*4*). Studying human infants, in the period when conceptual self is believed to emerge, therefore offers a unique opportunity to identify the foundations upon which an explicit self-representation might be built, shedding light on the structure and mechanisms of this characteristically human capacity.

To date, much of developmental work concerning self-related processes has focused on the bodily self (*5*, *6*), with empirical data providing evidence for how infants perceive and learn to distinguish their own bodies from those of others (*5*). Another line of theory and investigation focused on the potential role of social interactions in scaffolding infants' developing self-representation (*3*, *7–9*). Indeed, documented cultural differences in the age of acquisition of a conceptual self-representation are pointing to the importance of different social and child-rearing practices in plausibly mediating this development (*10*, *11*). However, while some have hypothesised a relationship between the two (*12*), the mechanisms through which more basic levels of self-perception might give rise to a reflective and conceptual self-representation remain unknown (*3*). Moreover, although interoceptive bodily signals may be the earliest available in development (*13*), a mature conceptual self-representation includes a crucial cognitive component: we monitor, control, and explicitly reflect on the activities of our own minds, not only our bodies. In adults, an individuals' online evaluation of own cognitive performance (i.e. confidence) can shape, and is in turn shaped by, the participants' broader self-beliefs (e.g. about own skills and abilities) (*14*, *15*). This mutual influence between cognitive self-monitoring and self-beliefs in adults thus suggests a promising yet unexplored route to investigating the building blocks of our reflective minds lies in exploring infants' early capacities for monitoring their own cognition.

Processes of cognitive self-monitoring unfold at multiple levels, from rapid neural signals that detect conflict or error (*16*), to evaluations of one's own perception, decisions, and informational needs (*17*), to fully explicit, conceptually mediated self-reflective capacities or metacognition (*18*, *19*). While explicit metacognition – the capacity to take one's mental states as the object of thought itself, making them reportable and shareable (*18*) – presupposes conceptual self-representation (*20*), other self-monitoring processes, such as internal error-monitoring, can be instantiated in simpler computational mechanisms, the output of which needs not to be consciously accessed (*21*) or meta-represented (*22*). Consistent with this, past decades of research have provided accumulating evidence for behaviours as well as neural correlates consistent with the presence of cognitive self-monitoring capacities in subjects considered not to possess a conceptual self-representation, namely non-human animals (*23*) and preverbal human

infants (*24*). However, although self-monitoring cognitive processes may precede conceptual self-representation in ontogeny as well as phylogeny, no data exists on whether these capacities are in fact developmentally related. A predictive link between the two would imply that self-monitoring processes not only emerge before self-representation, but that they potentially provide input or constitute a building block of the developing concept of self. To address this question, we follow a group of human infants across the critical period of development, when conceptual self is believed to emerge, to investigate whether infants' early implicit self-monitoring capacities, namely internal error-monitoring, developmentally predict the acquisition of a conceptual self-representation in the same infants.

In developmental research, the hallmark of having acquired a conceptual representation of the self is indicated by infants' success on the mirror self-recognition (MSR) test. Self-representation is attributed to a child if, when placed in front of a mirror while bearing a mark previously placed on their face unbeknownst to them, the child reaches for the mark on their own body (*25*). Although some have questioned the relevance of MSR for conceptual self-knowledge (e.g., *26*), the success on the MSR has been associated with a host of other co-emerging behaviours indicative of self-representation, such as infants using personal pronouns (*27*), exhibiting self-conscious emotions like embarrassment (*28*), and showing a self-reference memory bias (*29*). Combined, the evidence therefore suggests that success on the MSR is a robust and meaningful proxy for the presence of a conceptual self-representation in infants. We capitalise on the fact that, in western cultures, approximately 50% of infants exhibit such evidence of self-representation by 18 months of age (*3*). Harnessing this variability, we investigate whether earlier mechanisms of cognitive self-monitoring predict which infants will evidence conceptual self-awareness amongst age-matched 18-month-olds.

Our primary measure of infants' cognitive self-monitoring is error-related negativity (ERN), a well-established electroencephalography (EEG) signature of internal error monitoring in adults. The ERN is a negative deflection in frontocentral brain activity originating from the anterior cingulate cortex and arising after an individual commits an error (*30*). It is believed to reflect a second-order, post-decisional mechanism, signalling a discrepancy between the ongoing (or just-performed) action and the intended action (*16*). Importantly, this comparison is internal and does not rely on external feedback (*16*). Moreover, ERN occurs only when participants are conscious of the action they ought to perform, and absent if stimuli are subliminal, or if subjects are ignorant of the correct option (*30*). In infants, one prior study has demonstrated the presence of an ERN in 12-month-olds, when infants made incorrect predictive eye-movements (*24*). Like in adults, the infant ERN was not found when the stimuli were presented subliminally and occurred prior to infants receiving external feedback. Here, we measure the ERN as an index of early self-monitoring capacities in 12-month-old infants and investigate its predictive relationship to later emerging conceptual self-representation in the same infants, 6 months later.

At their first visit, 12-month-old infants took part in a novel, infant-controlled, decision-making task, based on match-to-sample paradigms (*31*, *32*; details in Methods). Infants interacted with a screen via an eye-tracker, which used gaze position to trigger different

parts of the display in response to infants' eye movements, in real time. This gave infants control of the task's timing and progression and enabled free exploration and decision-making. In each test trial, infants were presented with 3 cards, 2 of which were initially facing up, and 1 facing down (Figure 1). The face-up cards each displayed a different image of items familiar to infants (animals, fruit, etc.), which infants were free to look at for up to 20s (*exploration phase*). Once infants gazed at the third (face-down) card, their gaze triggered the reveal of the card's image, while simultaneously hiding the former two. The image on the third card always matched the image on one of the initially seen cards (now facing down) and was presented for up to 10s (*decision phase*). During this time, infants could gaze back at either of the former two cards and reveal their image again (*feedback phase*), thereby making a match (correct choice) or not (incorrect choice). Correct choices led to a reinforcing audio-visual animation (*reward*), while incorrect choices caused the cards to silently fade out (*no reward*). Crucially, for infants to reveal the image of a selected card, they had to gaze at it for cumulative 500ms. This allowed us to extract infants' neural activity (ERN) in the period after they had made their decision (their gaze landed on a face-down card) but before they had received any external feedback on their choice (image revealed).

If infants internally monitor their own decisions and have the capacity of detecting a discrepancy between the choice they had made, and the one they should have made, an ERN should be observed: frontocentral neural activity (aligned to the moment infants' gaze landed on the selected card) was expected to be more negative following incorrect as compared to correct choices (directional one-tailed hypothesis). Data from *N*=58 infants indeed revealed an ERN: cluster-based permutation analysis revealed two time periods in which the signals for correct and incorrect choices significantly diverged, an early peak at 33-172ms post-choice (p=0.031), and a late peak at 332-547ms (p=0.013). Both peaks fully matched the ERN previously observed in infants (*24*), while only the early peak matched the timing and polarity of the adult ERN, therefore the remainder of the analyses focuses on the first peak. To further validate that infants could distinguish between correct and incorrect choices, we additionally analysed infants' neural responses to external feedback (aligned to the reveal of the selected card's image). The same procedure revealed a feedback related negativity effect (FRN) over the same scalp region (two time periods were identified, early peak: 46-298ms, p=0.017; later peak: 457-930ms, p=0.001, the latter of which matched the timing of the adult FRN and was thus retained for further analyses). Combined, the observed neural responses suggest that 12-month-olds distinguished between correct and incorrect choices they had made, based both on external feedback (FRN) and, crucially, even prior to receiving the feedback (ERN).

Having established that 12-month-olds exhibit internal error-monitoring in our paradigm, we investigated whether this capacity predicts the later emergence of self-representation, by testing the same infants 6 months later, using the MSR task. We used a binary score on the MSR as the dependent outcome (success: *S* group, N=33; or not: *NS* group, N=20), and the magnitude of the ERN as the predictor (peak amplitude of frontocentral potentials for correct minus incorrect choices). A significant logistic regression model emerged (p=0.031): ERN at 12 months was found to correctly predict infants' self-representation 6 months later with 69.8% accuracy, correctly classifying

90.9% of *S* and 35.0% *NS* infants. Importantly, while the target behaviour in the MSR task (touching the mark on own face) is taken as evidence of self-representation, the absence of this behaviour is inconclusive. That is, there could be many reasons for a child not exhibiting the expected behaviour (e.g. shyness), and therefore the NS group is expected to be more heterogenous, thereby resulting in low classification accuracy for this group. To account for possible differences in infants' general development, a further stepwise model was conducted, which additionally included infants' age and a standardised measure of cognitive development (ASQ, see Methods). Both variables were rejected as relevant predictors of MSR, and anecdotal evidence was found for no group differences in either age or general cognitive development (BF_{01}=2.446 and BF_{01}=2.455, respectively). Our results thus indicate that the more sensitivity an infant shows in internally monitoring their own mistakes at 12 months (the stronger the ERN response), the more likely they are to have acquired a conceptual self-representation by the age of 18 months (as indicated by mirror self-recognition), irrespective of their age or cognitive development. Finally, no evidence was found for a predictive effect of the FRN on later self-representation (p=0.946), suggesting it is specifically infants' internal (but not external) error-monitoring at 12 months that is associated with infants' later emergence of self-representation at 18 months.

To further characterize the found developmental relationship, we analyse infants' early self-monitoring capacities separately for infants that showed self-representation at 18 months (*S* group) and those that did not (*NS* group). Peak analysis of the ERN response (37-172ms) revealed very strong evidence of internal error-monitoring in the *S* group only (BF_{10}=230.327). In contrast, moderate evidence for the absence of an ERN was found in the *NS* group (BF_{01}=4.188). Importantly, this could not be explained by differences in data quality or group size, as moderate evidence for the FRN effect was found in both groups (*S*: BF_{10}=4.223; *NS*: BF_{10}=6.147) with no group difference (BF_{01}=3.526). These results thus indicate that both groups of infants distinguished between correct and incorrect choices based on feedback, but only the *S* group additionally monitored their own decisions internally.

Finally, we explored whether internal error-monitoring not only predicts infants' self-representation 6 months later, but also associates with concurrent behavioural adjustments on task. Recorded as part of the same experiment at 12 moths of age, infants eyetracking data was analysed for evidence of two types of post-error adjustments. Firstly, the *decision phase* was analysed for evidence of post-error slowing, the commonly observed effect of increased decision times following the commitment of an error (*33*). Secondly, task-specific strategic adjustments were investigated in the *exploration phase,* namely increased post-error information gathering in the form of fixations on, and comparative looks between, the initially shown images. Comparing trials following an error to trials following a successful match, we found that only infants who evidenced internal self-monitoring and later self-representation (*S* group) showed evidence of behavioural adjustments: moderate evidence of post-error decision slowing (BF_{10}=5.107), strong evidence of increased post-error fixations (BF_{10}=10.467), and anecdotal evidence of increased post-error comparative looks (BF_{10}=1.396). This was not the case in the *NS* group: moderate evidence was found for the absence of decision slowing (BF_{01}=6.643), and anecdotal evidence for no increase in fixations and

comparative looks following errors (BF_{01}=1.078 and BF_{01}=2.242, respectively). Furthermore, in the *S* group, moderate evidence was found for a positive relationship between ERN magnitude and performance (average local derivative of infants' overall accuracy score, see Methods), suggesting the capacity to internally monitor own errors played a functional role in infants' overall performance on the task (R=0.420, BF_{10}=6.698). In contrast, moderate evidence for the absence of a relationship between ERN and performance was found in the *NS* group (R=-0.229, BF_{01}=6.647). Interestingly, despite the relationship between ERN and performance in *S* group, the absence of ERN and of behavioural adjustments in the *NS* group was not found to result in worse performance overall (anecdotal evidence for no group difference, BF_{01}=2.380). This suggests that comparable performance may be achieved via routes other than internal error monitoring and behavioural adjustments probed in this study. Taken together, our task's neural and behavioural measures provide converging evidence suggesting that, already at 12 months of age, infants' internal error-monitoring can aid performance and can be utilised for post-error behavioural adaptations. Crucially, this was only the case for infants who also exhibited evidence of self-awareness at 18 months.

In summary, our findings revealed that preverbal human infants' capacity to internally monitor their own errors developmentally predicts the emergence of their conceptual self-representation 6 months later. This relationship was specific to internal error-monitoring, and not external feedback processing, and could not be explained by infants' age or general cognitive development. Moreover, the group of infants, who evidenced internal error-monitoring, also demonstrated behavioural adjustments following errors. Combined, this data indicates that error-detection, if available, can be utilised for adaptive information-seeking and decision-making as early as 12 months of age. Finally, the magnitude of infants' response to errors correlated with their performance on task, but did not result in better overall performance, when compared to infants who did not evidence internal error-monitoring nor later self-representation. This suggests that the found relationship between early error-monitoring and later conceptual self-representation is unlikely a reflection of general differences in infants' abilities but is specific to infants' developing capacities of self-reflection.

The progression of self-reflective capacities from earlier self-monitoring to later self-representation, evidenced here, had already been hinted at by comparative research. For example, the information-seeking behaviours of capuchin monkeys (*34*), dogs (*35*), and even rats (*36*), have been argued to be driven by internal monitoring of cognitive states such as uncertainty. Moreover, although research on ERN in animals is virtually absent, one study recently evidenced single-neuron intracranial correlates of the ERN in capuchin monkeys (*37*). In contrast, with the exception of humans and great apes, robust evidence for self-representation in the animal kingdom is sparse and highly contested (*38*). Self-reflective capacities might therefore be another example of ontogeny recapitulating phylogeny (*39*), whereby both in evolution and in development, conceptual self-representation is preceded by earlier cognitive self-monitoring mechanisms. Given the seemingly similar initial capacities of cognitive self-monitoring, what explains the fact that humans develop an explicit conceptual representation of themselves, while other animals appear not to? It is possible that what seems equivalent on a behavioural level is in fact driven by different cognitive processes. Indeed, it has been argued that

behavioural adjustments seen in animals, such as seeking information in face of uncertainty, can be driven by purely first-order representations, namely by computations of the noise in the signal, without an additional, second-order metarepresentation of uncertainty (*40*). While the same argument can (and has been (*41*)) levied against attributing second-order representations to human infants, the results of the current study show that internal error-monitoring not only precedes but is in fact developmentally predictive of the emergence of infants' self-representation. This link thus offers strong support for the proposal that, at least in humans, internal error-monitoring and the corresponding behavioural adjustments index rudimentary forms of metacognition (*42*), and as such are indeed part of self-reflective capacities, and are not merely driven by first-order computations. Moreover, although current data cannot speak to this, the found developmental link opens the yet unexplored possibility that internal error-monitoring may be functionally involved in the process of building a concept of self.

One possibility for how error-monitoring may contribute to the building of a conceptual self-representation is that early availability of internal error signals can lead to better tuned metacognitive feelings. Defined as affective responses stemming from monitoring the successes and failures of own cognitive actions (*17*), metacognitive feelings can initiate epistemic behaviours, such as targeted information sampling. Such error-guided information seeking could facilitate the gradual building of conceptual representations in general, including the concept of self specifically. Alternatively, a more direct route from errors to self, but similarly grounded in error-guided information seeking, could be found in proposals emphasising one's tracking of own learning progress as the mechanism determining which activities an individual will be motivated to pursue (*43*). According to this account, information related to, and generated by, the self is particularly informative and allows for comparatively rapid learning progress, thereby driving self-related information seeking, and consequently the early emergence of a category of self (*44*). Although learning progress account does not distinguish between internally generated error signals and external feedback, it is plausible that awareness of errors as one's own would facilitate meta-learning, that is making use of one's prior experiences to learn how to learn better in the future (*45*). The behavioural adjustments following errors, seen in infants who evidenced internal error monitoring in our study, may therefore reflect infants meta-learning about their own information-seeking and decision-making strategies. As such, error tracking and internal evaluations of own learning may themselves become an integral part of the conceptual understanding of the self as an epistemic agent, as previously hinted at by adult research (*14*, *15*).

Much like it has been argued that interoceptive signals ought to be brought to awareness for them to provide a bridge between a biological and a conceptual self in development (*3*), it seems logically necessary that error-detection signals must (eventually) be consciously accessed, if error-monitoring is to be utilised for, or integrated into, a conceptual self-representation. Although the mechanisms enabling conscious access are believed to be already in place during the first year of life (*46*), it is unclear whether error detection in infants occurs on a conscious level, given that the ERN has occasionally been recorded in absence of a subjective report of having made an error even in adults (*21*). One challenge of specifying the process and mechanisms of building a conceptual self-representation thus lies in specifying the mechanism of conscious

access to mental representations itself. Both adult as well as developmental theories have put forward proposals as to how the mind may generate conscious experience (*47*), or turn information that is represented in the mind into progressively more explicit information that is accessible to the mind (*48*), however, empirical evidence is yet to provide conclusive data regarding the mechanisms supporting such transformations.

Finally, as the current evidence demonstrates, error-detection can lead to appropriate online adjustments in information-seeking and decision-making and can facilitate task performance even prior to any evidence of a conceptual self-representation. This raises another open question, namely, if cognitive control and adaptive learning can operate in absence of a conceptual self, why would humans nonetheless develop a conceptual self-representation or explicit metacognition? In addressing this question, some theoretical proposals have focused on plausible evolutionary pressures for evolving explicit reflective capacities and proposed the human-specific needs to reason about others' minds (*49*) and to coordinate mental content between multiple individuals (*18*) as the best candidates. For example, it is argued that by sharing their minds, humans were able to create and sustain the social structures and cumulative culture, characteristic of our species (*18*). Other, mainly developmental, theories have focused on the human-specific input that may trigger a conceptual self to develop and have likewise emphasised social interactions, identifying direct eye gaze (*3*) and turn-taking contingencies (*7*) as plausibly involved in the gradual emergence of a conceptual self. Interestingly, empirical evidence from children suggests that individuals' response to errors (as reflected in the magnitude of the ERN) may similarly be modulated by different aspects of social contexts (*50*). This raises the intriguing new possibility that one way in which early social input may affect the development of a self-representation is through modulating infants' detection and sensitivity to own errors.

The aim of the current study was to bring developmental data to bear on one of the foundational and enduring questions about human nature: how does our mind come to reflect upon itself? Our results revealed a developmental bridge between implicit error-monitoring and explicit self-representation, providing empirical traction on questions that have animated philosophers and cognitive scientists since Descartes' *cogito*. Our data strongly supports the proposal that preverbal infants already employ a rudimentary form of metacognition (*42*) and that this capacity predicts, and plausibly builds, their later conceptual representation of themselves as distinct agents. What was once an epistemic puzzle may now be approached as a testable developmental mechanism: from monitoring of errors to the emerging self, "I err, therefore I am."

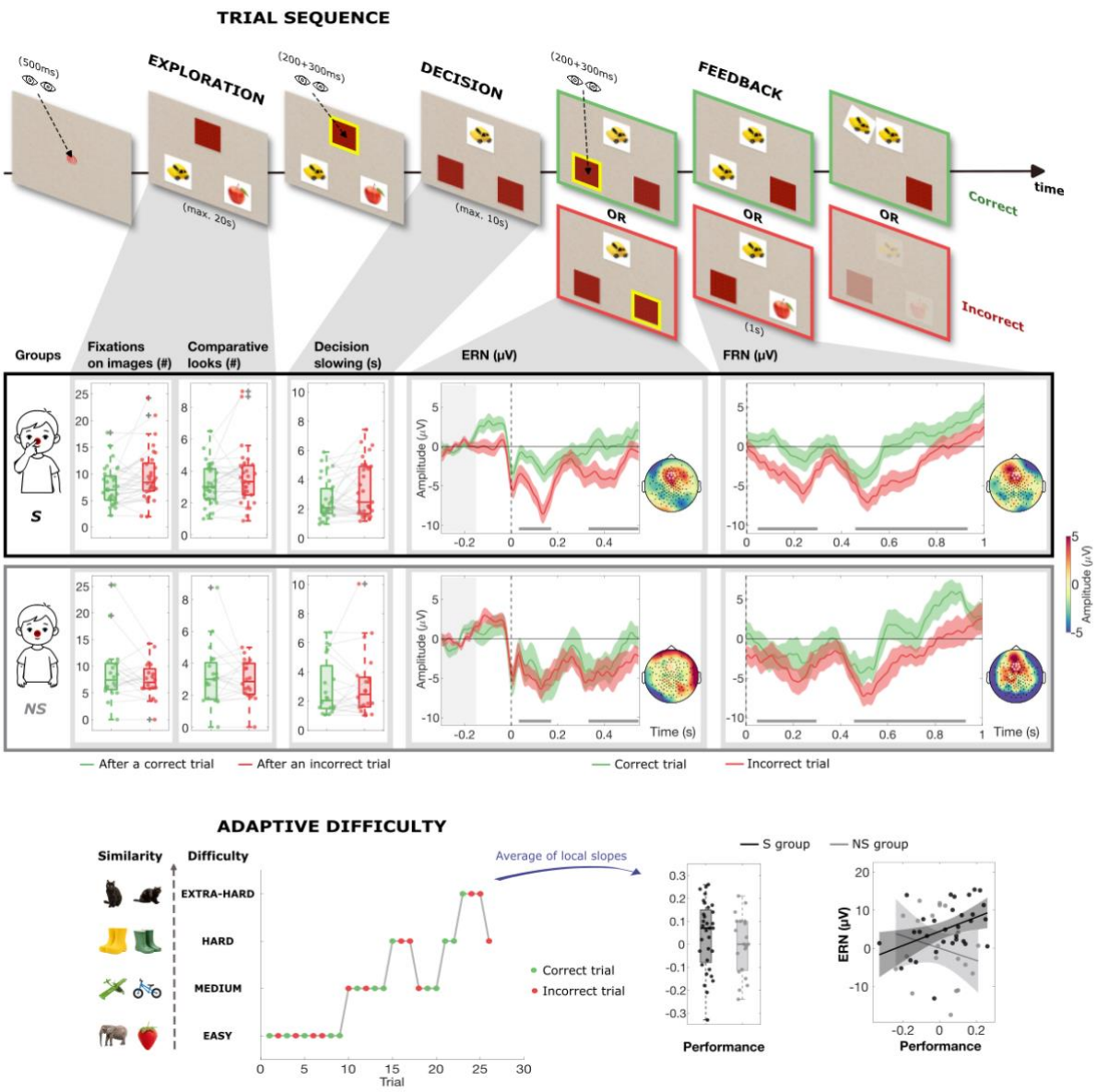

Figure 1. Task design and results. **Trial sequence (top):** screenshots of the different trial steps are depicted along the 3 phases (exploration, decision, feedback) with eye symbols representing the gaze-contingent parts of the trials. Green and red highlights around screenshots represent the sequence for correct or incorrect choices respectively. Results for the different eye-tracking and EEG measures at 12 months are presented under the trial phase at which they were extracted, for the two MSR groups separately: the group who showed evidence of self-representation at 18 months (*S*) in the black box; and the group who did not (*NS*) in the grey box. Correct (or following correct) trials are depicted in green, while incorrect (or following incorrect) trials are depicted in red. Eye tracking data (left) is presented in boxplots. Boxes represent the 25 to 75 percentiles, lines depict the median, and whiskers represent the range of included data points. Individual data points are represented by dots, and outliers are depicted by crosses. Average ERPs (right) are represented with lines, with shading representing the SEM. Solid lines at the bottom of the plots represent the time windows identified as significant by the time-cluster analysis, while topographies of the signal within the time window of interest for each group and signal (ERN or FRN) are plotted to the right of each plot. **Adaptive difficulty (bottom):** an example of a participants' sequence of trials (x axis) and the corresponding difficulty level of the trials (y axis) is depicted. Accuracy of the participants' responses is colour-coded (correct trials in green, incorrect in red) and example pairs of stimuli along each level of difficulty are shown on the side (moving up in difficulty after two consecutive correct choices, down after 2 consecutive incorrect). The groups' performance (average local slope in the participant's difficulty level as a function of trial number) is depicted for the *S* group in black and the *NS* group in grey, together with the groups' correlation between ERN amplitude and performance.

References

1. S. M. . Fleming, Know thyself : the science of self-awareness. 290 (2021).
2. J. Proust, Metacognition. *Philos. Compass* **5**, 989–998 (2010).
3. V. Southgate, The Origins and Emergence of Self-Representation. *Annu. Rev. Dev. Psychol.* **6**, 109–131 (2024).
4. P. Rochat, Five levels of self-awareness as they unfold early in life. *Conscious. Cogn.* **12**, 717–731 (2003).
5. M. L. Filippetti, S. Lloyd-Fox, M. R. Longo, T. Farroni, M. H. Johnson, Neural Mechanisms of Body Awareness in Infants. *Cerebral Cortex* **25**, 3779–3787 (2015).
6. J. Yang, M. K. Yamaguchi, A. J. Bremner, Development of body representations in human infancy. *Philosophy & cultural embodiment* **1** (2021).
7. G. Gergely, J. S. Watson, The social biofeedback theory of parental affect-mirroring: The development of emotional self-awareness and self-control in infancy. *International Journal of Psycho-Analysis* **77**, 1181–1212 (1996).
8. A. Fotopoulou, M. Tsakiris, Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis* **19**, 3–28 (2017).
9. M. L. Filippetti, Being in Tune With Your Body: The Emergence of Interoceptive Processing Through Caregiver–Infant Feeding Interactions. *Child Dev. Perspect.* **15**, 182–188 (2021).
10. H. Keller, R. Yovsi, J. Borke, J. Kärtner, H. Jensen, Z. Papaligoura, Developmental Consequences of Early Parenting Experiences: Self-Recognition and Self-Regulation in Three Cultural Communities. *Child Dev.* **75**, 1745–1760 (2004).
11. S. Cebioğlu, T. Broesch, Explaining cross-cultural variation in mirror self-recognition: New insights into the ontogeny of objective self-awareness. *Dev. Psychol.* **57**, 625–638 (2021).
12. C. E. Palmer, M. Tsakiris, Going at the heart of social cognition: is there a role for interoception in self-other distinction? *Curr. Opin. Psychol.* **24**, 21–26 (2018).
13. A. W. Corcoran, K. Perrykkad, D. Feuerriegel, J. E. Robinson, Body as First Teacher: The Role of Rhythmic Visceral Dynamics in Early Cognitive Development. *Perspectives on Psychological Science* **20**, 45–75 (2025).
14. M. Rouault, P. Dayan, S. M. Fleming, Forming global estimates of self-performance from local confidence. *Nat. Commun.* **10:1141** (2019).
15. M. Rouault, S. M. Fleming, Formation of global self-beliefs in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 27268–27276 (2020).
16. S. Dehaene, The Error-Related Negativity, Self-Monitoring, and Consciousness. *Perspect. Psychol. Sci.* **13**, 161–165 (2018).
17. L. Goupil, J. Proust, Curiosity as a metacognitive feeling. *Cognition* **231**, 105325 (2023).
18. N. Shea, A. Boldt, D. Bang, N. Yeung, C. Heyes, C. D. Frith, Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18**, 186–193 (2014).
19. J. H. Flavell, Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* **34**, 906–911 (1979).
20. L. D. Frazier, B. L. Schwartz, J. Metcalfe, The MAPS model of self-regulation: Integrating metacognition, agency, and possible selves. *Metacognition and Learning 2021 16:2* **16**, 297–318 (2021).
21. S. Nieuwenhuis, K. R. Ridderinkhof, J. Blom, G. P. H. Band, A. Kok, Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology* **38**, 752–760 (2001).
22. T. Gliga, V. Southgate, Metacognition: Pre-verbal Infants Adapt Their Behaviour to Their Knowledge States. *Current Biology* **26**, R1191–R1193 (2016).
23. B. L. Schwartz, A. Pournaghdali, K. L. Hess, Comparative Approaches to the Natural Ecology of Metacognition. *Comp. Cogn. Behav. Rev.* **18**, 79 (2023).
24. L. Goupil, S. Kouider, Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology* **26**, 3038–3045 (2016).
25. G. G. Gallup, Self-awareness and the evolution of social intelligence. *Behavioural Processes* **42**, 239–247 (1998).
26. C. M. Heyes, Reflections on self-recognition in primates. *Anim. Behav.* **47**, 909–919 (1994).
27. M. Lewis, D. Ramsay, Development of Self-Recognition, Personal Pronoun Use, and Pretend Play During the 2nd Year. *Child Dev.* **75**, 1821–1831 (2004).
28. M. Lewis, M. W. Sullivan, C. Stanger, M. Weiss, Self Development and Self-Conscious Emotions. *Child Dev.* **60**, 146 (1989).
29. C. Grosse Wiesmann, K. Rothmaler, E. Hasan, K. Habdank, C. Yang, E. Yeung, V. Southgate, The self-reference memory bias is preceded by an other-reference bias in infancy. *Nature Communications 2025 16:1* **16**, 6311- (2025).
30. L. Charles, F. Van Opstal, S. Marti, S. Dehaene, Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* **73**, 80–94 (2013).
31. J. R. Hochmann, S. Mody, S. Carey, Infants' representations of same and different in match- and non-match-to-sample. *Cogn. Psychol.* **86**, 87–111 (2016).

32. Z. Kaldy, S. B. Guillory, E. Blaser, Delayed Match Retrieval: a novel anticipation-based visual working memory paradigm. *Dev. Sci.* **19**, 892–900 (2016).
33. P. M. Rabbitt, Errors and error correction in choice-response tasks. *J. Exp. Psychol.* **71**, 264–272 (1966).
34. M. J. Beran, J. D. Smith, Information seeking by rhesus monkeys (Macaca mulatta) and capuchin monkeys (Cebus apella). *Cognition* **120**, 90–105 (2011).
35. J. Belger, J. Bräuer, Metacognition in dogs: Do dogs know they could be wrong? *Learn. Behav.* **46**, 398–413 (2018).
36. S. Yuki, Y. Sakurai, D. Yanagihara, Rats adaptively seek information to accommodate a lack of information. *Scientific Reports 2023 13:1* **13**, 14417- (2023).
37. Z. Fu, A. Sajad, S. P. Errington, J. D. Schall, U. Rutishauser, Neurophysiological mechanisms of error monitoring in human and non-human primates. *Nature Reviews Neuroscience 2023 24:3* **24**, 153–172 (2023).
38. G. G. Gallup, J. R. Anderson, Self-recognition in animals: Where do we stand 50 years later? Lessons from cleaner wrasse and other species. *Psychology of Consciousness: Theory Research, and Practice* **7**, 46–58 (2020).
39. S. A. Heldstab, K. Isler, C. Schuppli, C. P. van Schaik, When ontogeny recapitulates phylogeny: Fixed neurodevelopmental sequence of manipulative skills among primates. *Sci. Adv.* **6** (2020).
40. P. Carruthers, Meta-cognition in animals: A skeptical look. *Mind Lang.* **23**, 58–89 (2008).
41. T. Gliga, V. Southgate, Metacognition: Pre-verbal Infants Adapt Their Behaviour to Their Knowledge States. *Current Biology* **26**, R1191–R1193 (2016).
42. L. Goupil, S. Kouider, Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Curr. Dir. Psychol. Sci.* **28**, 403–408 (2019).
43. J. Gottlieb, P.-Y. Oudeyer, M. Lopes, A. Baranes, Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* **17**, 585–93 (2013).
44. F. Kaplan, P. Y. Oudeyer, "The progress drive hypothesis: An interpretation of early imitation" in *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions* (Cambridge University Press, 2007; http://pyoudeyer.com/kaplan-05a.pdf), pp. 361–378.
45. F. Poli, T. Ghilardi, R. B. Mars, M. Hinne, S. Hunnius, Eight-Month-Old Infants Meta-Learn by Downweighting Irrelevant Evidence. *Open Mind* **7**, 141–155 (2023).
46. S. Kouider, C. Stahlhut, S. V. Gelskov, L. S. Barbosa, M. Dutat, V. De Gardelle, A. Christophe, S. Dehaene, G. Dehaene-Lambertz, A Neural Marker of Perceptual Consciousness in Infants. *Science (1979).* **340**, 376–380 (2013).
47. O. Ferrante, U. Gorska-Klimowska, S. Henin, R. Hirschhorn, A. Khalaf, A. Lepauvre, L. Liu, D. Richter, Y. Vidal, N. Bonacchi, T. Brown, P. Sripad, M. Armendariz, K. Bendtz, T. Ghafari, D. Hetenyi, J. Jeschke, C. Kozma, D. R. Mazumder, S. Montenegro, A. Seedat, A. Sharafeldin, S. Yang, S. Baillet, D. J. Chalmers, R. M. Cichy, F. Fallon, T. I. Panagiotaropoulos, H. Blumenfeld, F. P. de Lange, S. Devore, O. Jensen, G. Kreiman, H. Luo, M. Boly, S. Dehaene, C. Koch, G. Tononi, M. Pitts, L. Mudrik, L. Melloni, Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature 2025 642:8066* **642**, 133–142 (2025).
48. A. Karmiloff-Smith, "Precis of Beyond modularity: A developmental perspective on cognitive science" (1994).
49. P. Carruthers, How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* **32**, 121–138 (2009).
50. J. C. García Alanis, T. E. Baker, M. Peper, M. L. Chavanon, Social context effects on error-related brain activity are dependent on interpersonal and achievement-related traits. *Scientific Reports 2019 9:1* **9**, 1728- (2019).