Developing a self-reflective mind: infants' error-detection at 12 months predicts self-awareness at 18 months

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 has been proposed as one of the key drivers of our progress as a species, yet the development of self-reflective capacities is unknown. Using a novel 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 neural response signature of error-tracking (the error-related negativity, ERN) when they make incorrect as opposed to correct choices and before receiving any feedback. Remarkably, the magnitude of the ERN at 12 months developmentally predicted the emergence of conceptual self-representation (mirror-self-recognition) in the same infants six months later. This relationship was specific to internal error-monitoring and not to external feedback processing, and could not be explained by infants' age or general cognitive development. Moreover, infants did not simply detect their errors: the group of infants who passed the mirror test at 18 months (*S*) also adapted their behaviour post-error at 12 months. Following mistakes, *S* infants slowed their decisions and increased their visual exploration, demonstrating not only rudimentary metacognitive monitoring but also metacognitive control at 1 year of age. These findings provide the first empirical evidence that self-awareness may be rooted in earlier self-monitoring abilities, and reveal how a self-reflective mind can emerge from implicit cognitive monitoring mechanisms within the first years of life.

Article main text

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*, *2*). Despite its centrality to human cognition, the mechanisms enabling our mind to reflect upon itself remain subject of much theoretical debate (*3*). One approach to tackling the enigma of human self-awareness, which has gained renewed interest in recent years (*4*), 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 (*5*). 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 might be built, shedding light on the structure and mechanisms of this characteristically human capacity.

Developmental work concerning self-related processes has largely focused on the bodily self (*6*, *7*) and the role of social interactions in scaffolding the developing self-representation (*4*, *8–10*). Empirical data has provided evidence for how infants perceive and learn to distinguish their own bodies from those of others (*6*), and documented cultural differences in the age of acquisition of a conceptual self-representation have pointed to the importance of different social practices in mediating this development (*11*, *12*). Yet, the mechanisms through which the basic levels of self-perception give rise to reflective self-awareness, and a self-concept, remain unknown (*4*). Moreover, although interoceptive bodily signals may be the earliest available in development, a mature conceptual self-representation includes a crucial cognitive component: the capacity to monitor, control, and explicitly reflect on our own cognitive activities. Therefore, a promising yet largely unexplored route to investigating the building blocks of our reflective minds lies in exploring infants' early capacities not for bodily interoception, but for cognitive self-monitoring.

Processes of cognitive self-monitoring unfold at multiple levels, from rapid neural signals that detect conflict or error (*13*), to evaluations of one's own perception, decisions, and informational needs (*14*), to fully explicit, conceptually mediated metacognition (*15*, *16*). While explicit metacognition – the capacity to take one's mental states as the object of thought itself, making them reportable and shareable – presupposes conceptual self-awareness (*17*), 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 (*18*) or meta-represented (*19*). 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 (*20*) and preverbal human infants (*21*). However, despite the evidence that self-monitoring cognitive processes may precede explicit self-awareness, no data exists on whether these capacities are developmentally related. To address this question, we follow a group of infants across the critical period of development from 12 to 18 months of age, to investigate whether infants' early implicit self-reflective capacities, namely internal error-monitoring, developmentally predict the acquisition of explicit self-representation in the same infants 6 months later.

In developmental research, the hallmark of having acquired a conceptual representation of the self is believed to be 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 (*22*). Although some have questioned the relevance of MSR for conceptual self-knowledge (e.g., *23*), 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 (*24*), exhibiting self-conscious emotions like embarrassment (*25*), and showing a self-reference memory bias (*26*) thereby making the MSR a robust and meaningful correlate 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 (*4*). 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 the 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 (*27*). 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 (*28*). Importantly, this comparison is internal and does not rely on external feedback (*13*). Moreover, ERN occurs only when subjects are conscious of the action they ought to perform, and absent if stimuli are subliminal or if subjects are ignorant of the correct option (*27*). In infants, one prior study has demonstrated the presence of an ERN in 12-month-olds, when infants made incorrect predictive eye-movements (*29*). Like in adults, the infant ERN was not found when the stimuli were presented subliminally and occurred prior to infants receiving external feedback.

At their first visit, 12-month-old infants took part in a novel, infant-controlled, decision-making task, based on match-to-sample paradigms (*30*, *31*; details in Methods). Infants interacted with a screen via an eye-tracker, which used infants' 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). 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 continuously gaze at it for 500ms. This allowed us to extract infants' neural activity 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) is 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 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 (*29*), 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). The same procedure revealed a feedback related negativity effect (FRN) over the same scalp region (two time periods 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, 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 internal error-monitoring at 12 months, 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 magnitude of the ERN as the predictor (peak amplitude of frontocentral potentials for correct minus incorrect choices). A significant logistic linear 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 (to note, the high-sensitivity/low-specificity of the model is explained by the fact that the *NS* group was negatively defined by the absence of a target behaviour). Furthermore, a stepwise model additionally including infants' age and general cognitive development rejected both variables as relevant predictors of MSR, and anecdotal evidence for no group difference was found for both variables (BF_{01}=2.446 and BF_{01}=2.455, respectively), suggesting that these results are unlikely to be explained by mere differences in infants' maturity. Our results thus indicate that the more sensitivity an infant shows in internally monitoring their own mistakes at 12 months, the more likely they are to have acquired a conceptual self-representation by the age of 18 months, providing the first evidence to directly link internal error-monitoring to the emergence of self-representation in development.

To further characterise the found developmental relationship, we compare and analyse 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 task understanding, 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, moderate). 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. Moreover, 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.

This developmental link strongly suggests that internal error monitoring may serve as one of the building blocks of the developing sense of self in infants. One possibility for how this may unfold in development is that early availability of internal error signals could lead to better tuned metacognitive feelings (*14*), which could in turn 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. Another possibility is that error signals are utilised in computations of learning progress, which has been proposed to be the driving force of infant (and adult) information seeking behaviours (*32*). According to this account, the self could emerge as a meaningful category distinguished from other elements in infants' environment based on its predictability and learning progress it affords (*33*). Critical to both proposals is that the capacity to detect errors should lead infants modulate their behaviour in response to errors. To probe whether infants show such adjustments in our task, infants' gaze behaviour was analysed for evidence of adjustments in their information-search and decision-making following errors.

We analysed the concurrently recorded eyetracking data for evidence of post-error slowing in the *decision phase* (*34*), and heightened information gathering in *exploration phase* (fixations on, and comparative looks between, the two images). Comparing trials following an error to trials following a successful match (directional one-tailed hypothesis), 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). Finally, moderate evidence was found for a positive relationship between ERN and performance (average local derivative of infants' overall accuracy score, see Methods) in the *S* group (R=0.420, BF_{10}=6.698), suggesting ERN played a functional role in infants' success on the task. 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). Importantly, despite evidence for the functional role of the ERN in aiding performance, the absence of ERN and behavioural adjustments in the *NS* group was not found to result in worse performance (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, with the *NS* group possibly relying on better memory or attentional capacities, which were not probed in this study. In sum, the converging evidence from our task's neural and behavioural indices is compatible with infant's internal error monitoring (ERN) being utilised for behavioural adaptation, and predictive of infants' performance on task. Crucially, this was only the case for infants who also exhibited evidence of self-awareness at 18 months.

In summary, the present study 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, evidence of infants' behavioural adjustments following errors in our study indicates that errors can be utilised for adaptive information seeking, as early as 12 months of age. Finally, infants' internal error-monitoring was found to correlate with their success on task, but did not result in better performance overall, suggesting that the relationship between early error-monitoring and later self-representation is not a reflection of superior cognitive abilities in general, but is specific to reading self-related signals.The fact that internal error-monitoring not only operates before the emergence of self-representation but also developmentally predicts its acquisition, provides the first evidence suggesting that conceptual self-awareness may at least in part be rooted in developmentally earlier cognitive-monitoring mechanisms. These results not only reveal a powerful learning mechanism, consistent with models of intrinsically motivated learning (*35*), but also hint at how human self-reflective capacities may gradually develop from implicit to explicit forms.

The developmental link evidenced here between early self-monitoring and later self-concept in human infants raises questions as to whether this link might also exist in evolution. Few species show clear evidence of spontaneous mirror self-recognition (e.g., *37*), and fewer studies exist to study metacognition in these species, but when conducted, these studies have repeatedly showed evidence for metacognition in mirror-passing species such as great apes (*38*, *39*) and dolphins (*40*). In contrast, species which clearly do not spontaneously pass the mirror test are still found to show signs of metacognition such as seeking information in the face of uncertainty, whether rats (*41*), dogs (*42*) or capuchin monkeys (*43*). Finally, although research on ERN in animals is virtually absent, one study recently evidenced single-neuron intracranial correlates of the ERN in the non-spontaneously mirror-passing capuchin monkeys (*44*). Taken together, animal research thus supports the idea that self-monitoring abilities as reflected by metacognitive behaviours or the ERN, could precede self-representation not just in development, as our study suggests, but also in evolution.

Consequently, one might argue that older philosophical accounts placing metacognitive reflection at the centre of the human experience, famously worded by Descartes as "I think, therefore I am", could lack precision for distinguishing the uniqueness of human self-awareness from non-human, metacognitive species. Theories positing high sociality as the driving factor explaining complex self-representation and mirror-passing both in animals (*45*) and infants (*46*) provide a helpful addendum. Sociality was not directly tested in the current study, but control questionnaire variables nonetheless bring in relevant data that could support this idea: higher scores specifically for communication and personal-social scales only were found in mirror-passers *S* vs. non-passers *NS* (respectively, at 18 months, BF_{10}=13.325 and BF_{10}=7.674; more in Supplementary Materials). Given humans' uniquely complex social interactions and communication, this could explain why only humans develop such a strong conceptual, explicit self. We believe that this collection of results has

broad and significant consequences, not only for our understanding of infant and human cognition, but also for animal and artificial cognition. On the latter, we propose that lack of sociability and second-order judgements, and poor metacognitive performance in current artificial intelligence (*47*), strongly advocate that artificial intelligence is currently still far from possessing the building blocks for self-awareness.

Figure

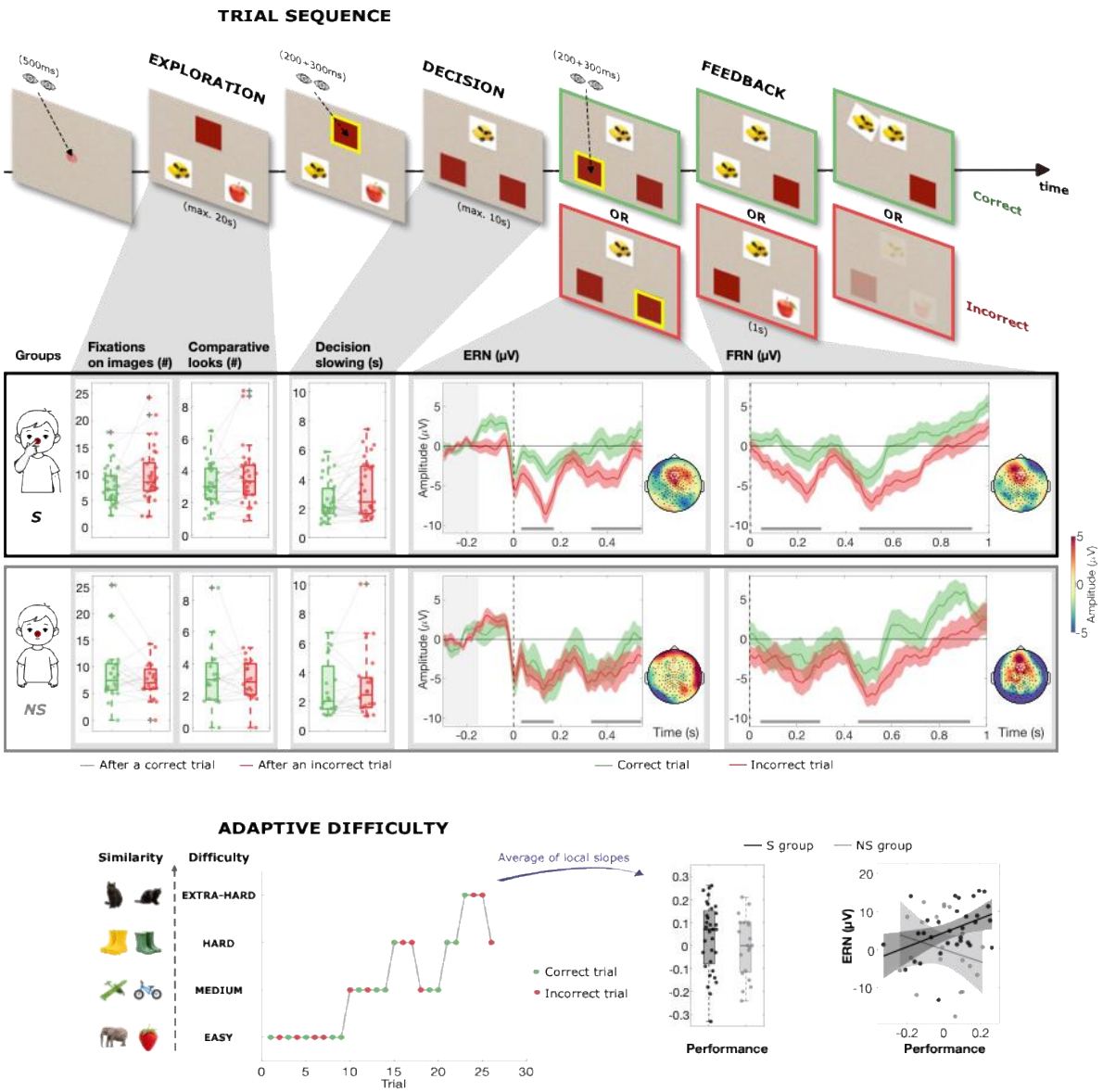


Figure: Task design and results. **Trial sequence:** screenshots of the different trial steps are depicted along the 3 phases (exploration, decision, feedback) with eyes connected to the screen with an arrow representing the gaze-contingent parts of the trials; green and red highlights around screenshots respectively represent the sequence for correct or incorrect choices. Results for the different eye-tracking and EEG measures at 12 months are presented under the trial stage at which they are extracted (screenshots and plots are linked by grey shadows) for the two MSR groups separately: the group who showed

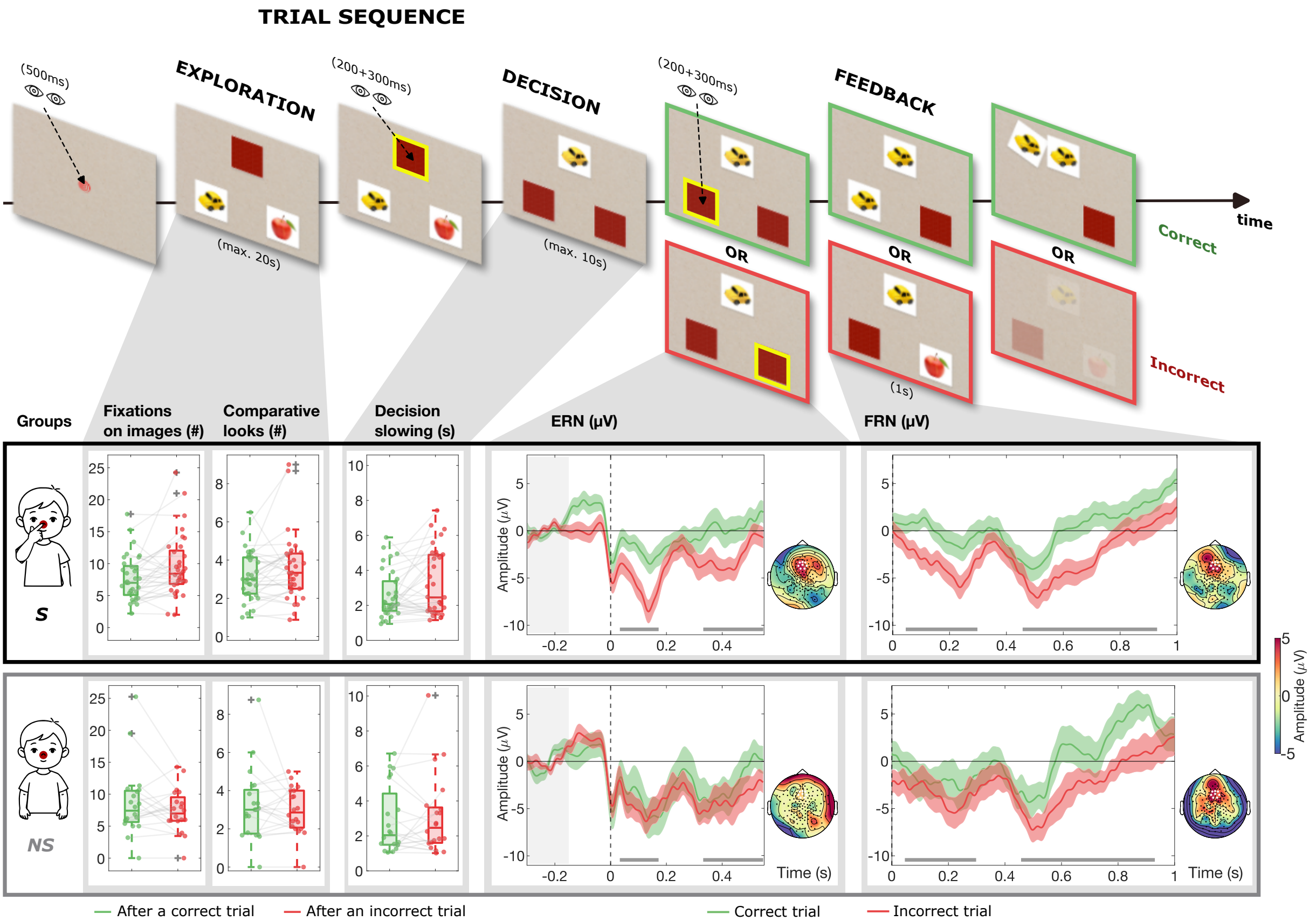
evidence of self-recognition at 18 months (*S*) in the black box and the group who did not (*NS*) in the grey box. Correct (or following correct) trials are depicted in green, while incorrect (or following incorrect) trials are depicted in red. Eye tracking data is presented in boxplots comprise a box representing the 25 to 75% percentiles split into two at the median line with whiskers representing the range of included data points, which are represented by dots, while outliers are depicted by crosses. For with shading representing the SEM; solid lines at the bottom of the plots represent the time windows identified as significant by the time-clustering 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:** an example plot of a participants' sequence of trials (x axis) and the corresponding difficulty level of the trials (y axis) are depicted, together with the accuracy of the participants' response (correct trials in green, incorrect in red) and example pairs of stimuli along each level of difficulty (moving up with two consecutive correct choices, 2 with 2 consecutive incorrect). The groups' performance (average local slope in the participant's difficulty level reached as a function of trial number) are depicted for the *S* group in black and the *NS* group in grey, together with the groups' correlation between ERN amplitude and performance.

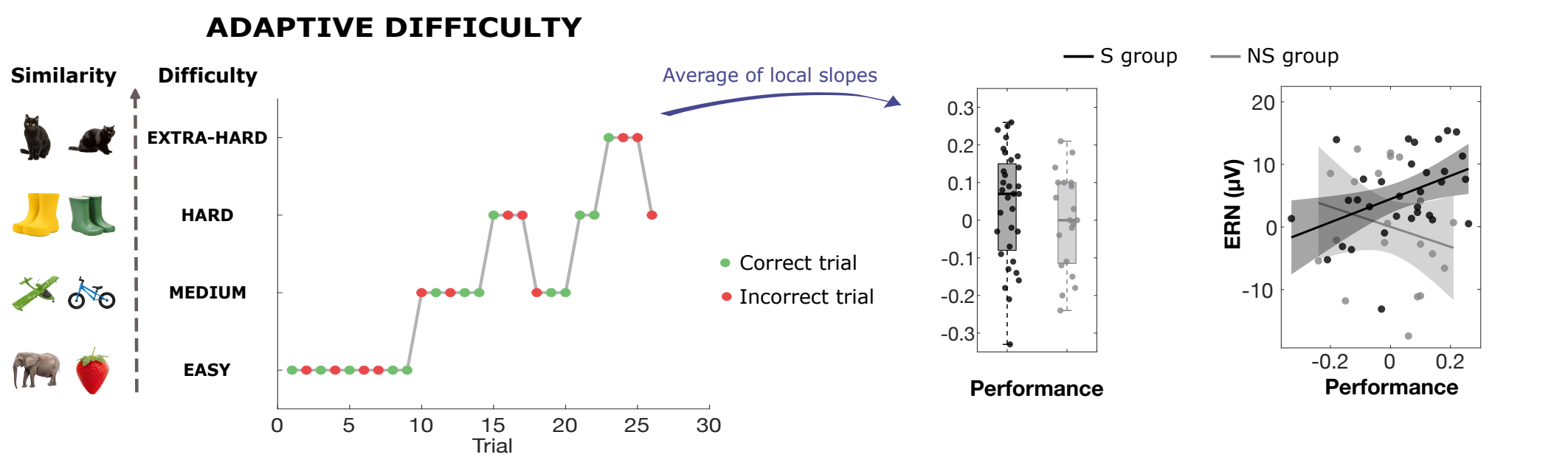
References

1. P. Rochat, The Ontogeny of Human Self-Consciousness. *Curr. Dir. Psychol. Sci.* 27, 345–350 (2018).
2. S. M. . Fleming, Know thyself : the science of self-awareness. 290 (2021).
3. 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).
4. V. Southgate, The Origins and Emergence of Self-Representation. *Annu. Rev. Dev. Psychol.* 6, 109–131 (2024).
5. P. Rochat, Five levels of self-awareness as they unfold early in life. *Conscious. Cogn.* 12, 717–731 (2003).
6. 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).
7. J. Yang, M. K. Yamaguchi, A. J. Bremner, Development of body representations in human infancy. *Philosophy & cultural embodiment* 1 (2021).
8. 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).
9. A. Fotopoulou, M. Tsakiris, Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28 (2017).

10. 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).
11. 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).
12. 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).
13. S. Dehaene, The Error-Related Negativity, Self-Monitoring, and Consciousness. *Perspect. Psychol. Sci.* 13, 161–165 (2018).
14. L. Goupil, J. Proust, Curiosity as a metacognitive feeling. *Cognition* 231, 105325 (2023).
15. 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).
16. J. H. Flavell, Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34, 906–911 (1979).
17. 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).
18. 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).
19. T. Gliga, V. Southgate, Metacognition: Pre-verbal Infants Adapt Their Behaviour to Their Knowledge States. *Current Biology* 26, R1191–R1193 (2016).
20. B. L. Schwartz, A. Pournaghdali, K. L. Hess, Comparative Approaches to the Natural Ecology of Metacognition. *Comp. Cogn. Behav. Rev.* 18, 79 (2023).
21. L. Goupil, S. Kouider, Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Curr. Dir. Psychol. Sci.* 28, 403–408 (2019).
22. G. G. Gallup, Self-awareness and the evolution of social intelligence. *Behavioural Processes* 42, 239–247 (1998).
23. C. M. Heyes, Reflections on self-recognition in primates. *Anim. Behav.* 47, 909–919 (1994).
24. M. Lewis, D. Ramsay, Development of Self-Recognition, Personal Pronoun Use, and Pretend Play During the 2nd Year. *Child Dev.* 75, 1821–1831 (2004).
25. M. Lewis, M. W. Sullivan, C. Stanger, M. Weiss, Self Development and Self-Conscious Emotions. *Child Dev.* 60, 146 (1989).
26. 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).
27. L. Charles, F. Van Opstal, S. Marti, S. Dehaene, Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* 73, 80–94 (2013).
28. L. Charles, J. R. King, S. Dehaene, Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli. *Journal of Neuroscience* 34, 1158–1170 (2014).
29. L. Goupil, S. Kouider, Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology* 26, 3038–3045 (2016).

30. 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).
31. Z. Kaldy, S. B. Guillory, E. Blaser, Delayed Match Retrieval: a novel anticipation-based visual working memory paradigm. *Dev. Sci.* 19, 892–900 (2016).
32. 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).
33. 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.
34. P. M. Rabbitt, Errors and error correction in choice-response tasks. *J. Exp. Psychol.* 71, 264–272 (1966).
35. F. B. M. de Waal, Fish, mirrors, and a gradualist perspective on self-awareness. *PLoS Biol.* 17, e3000112 (2019).
36. J. Call, M. Carpenter, Do apes and children know what they have seen? *Anim. Cogn.* 3, 207–220 (2001).
37. M. Tomasello, Social cognition and metacognition in great apes: a theory. *Anim. Cogn.* 26, 25–35 (2023).
38. D. Browne, Do dolphins know their own minds? *Biol. Philos.* 19, 633–653 (2004).
39. S. Yuki, Y. Sakurai, D. Yanagihara, Rats adaptively seek information to accommodate a lack of information. *Scientific Reports 2023 13:1* 13, 14417- (2023).
40. J. Belger, J. Bräuer, Metacognition in dogs: Do dogs know they could be wrong? *Learn. Behav.* 46, 398–413 (2018).
41. M. J. Beran, J. D. Smith, Information seeking by rhesus monkeys (Macaca mulatta) and capuchin monkeys (Cebus apella). *Cognition* 120, 90–105 (2011).
42. 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).
43. Y. Lei, Sociality and self-awareness in animals. *Front. Psychol.* 13, 1065638 (2023).
44. P. Rochat, Others in mind: Social origins of self-consciousness. *Others in Mind: Social Origins of Self-Consciousness*, 1–253 (2009).
45. B. Conway-Smith, R. L. West, Toward Autonomy: Metacognitive Learning for Enhanced AI Performance. *Proceedings of the AAAI Symposium Series* 3, 545–546 (2024).


TRIAL SEQUENCE
(500ms)
EXPLORATION
(max. 20s)
(200+300ms)
DECISION
(max. 10s)
(200+300ms)
FEEDBACK
OR
OR
OR
(1s)
time
Correct
Incorrect
Groups
Fixations on images (#)
Comparative looks (#)
Decision slowing (s)
ERN (μV)
FRN (μV)
S
NS
Amplitude (μV)
Time (s)
After a correct trial
After an incorrect trial
Correct trial
Incorrect trial


ADAPTIVE DIFFICULTY
Similarity
Difficulty
EXTRA-HARD
HARD
MEDIUM
EASY
Average of local slopes
Correct trial
Incorrect trial
Trial
S group
NS group
Performance
ERN (μV)
Performance