Neural correlates of performance monitoring vary as a function of competition between automatic and controlled processes: An ERP study

Nassim Elimari, Gilles Lafargue

Neural correlates of performance monitoring vary as a function of competition between automatic and controlled processes: An ERP study

Nassim Elimari[1], Gilles Lafargue[2] *

[1 2] *Université de Reims Champagne Ardenne, C2S, EA 6291, France.*

Presented here is the unedited draft copy of "*Neural correlates of performance monitoring vary as a function of competition between automatic and controlled processes: An ERP study*" originally published in *Consciousness and Cognition*

Citation

Elimari, N., & Lafargue, G. (2023). Neural correlates of performance monitoring vary as a function of competition between automatic and controlled processes: An ERP study. *Consciousness and Cognition*, *110*, 103505.

The Version of Record of this article, as published and maintained by the publisher, is available online at: https://doi.org/10.1016/j.concog.2023.103505

**Correspondance: Gilles Lafargue, Université de Reims Champagne-Ardenne, Laboratoire C2S, B.P. 30, 57 rue Pierre Taittinger, REIMS Cedex 51171. E-mail address: gilles.lafargue@univ-reims.fr*

Abstract

Dual process theories of attitude formation propose that an evolutionary old associative system automatically generates subjective judgments by processing mere spatiotemporal contiguity between paired objects, subjects, or events. These judgments can potentially contradict our well-reasoned evaluations and highjack decisional or behavioral outcomes. Contrary to this perspective, other models stress the exclusive work of a single propositional system that consciously process co-occurrences between environmental cues and produce propositions, i.e., mental statements that capture the specific manner through which stimuli are linked. We constructed an experiment on the premise that it would be possible, if the associative system does produce attitudes in a parallel non-conscious fashion, to condition two mutually exclusive attitudes (one implicit, the other explicit) toward a same stimulus. Through explicit ratings, inhibition performance, and neural correlates of performance monitoring, we assessed whether there was a discrepancy between stimuli that were conditioned with (1) the two systems working in harmony (i.e., producing congruent attitudes), or (2) the two systems working in competition (i.e., producing incongruent attitudes). Compared with congruent stimuli, incongruent stimuli consistently elicited more neutral liking scores, higher response times and error rates, as well as a diminished amplitudes in two well-studied neural correlates of automatic error detection (i.e., error-related negativity) and conscious appraisal of error commission (i.e., error-related positivity). Our findings are discussed in the light of evolutionary psychology, dual-process theories of attitude formation and theoretical frameworks on the functional significance of error-related neural markers.

Keywords: Dual-Process Theory, Attitude Formation, Valence, Error-related Negativity, Awareness, Evolutionary Psychology, Anterior Cingulate Cortex

1. INTRODUCTION

From an evolutionary standpoint, the adaptive benefits of associative learning (AL, defined as the behavioral change of an organism as a result of the processing of new relations between cues, e.g., Abramson, 1994; Ginsburg & Jablonka, 2010, Shanks, 1995) are so straightforward that they can only be described through a rather mundane statement: to survive in a dynamic and ever-changing world, organisms must flexibly adapt to ecological variations by forming and adjusting evaluations about environmental cues. As expected from such a fundamental adaptation, AL is evolutionary old and present in organisms with much less complex brains than ours, such as pigeons (e.g., George & Pearce, 1999; MacKintosh & Little, 1969; Skinner, 1948), rats (e.g., Garcia et al., 1968), mollusks (e.g., Hawkins et al., 1989), flatworms (e.g., Prados et al., 2013), or nematodes (e.g., Ardiel & Rankin, 2010). It has been proposed that AL emerged in early bilaterians and has been a leading factor driving the Cambrian explosion (Ginsburg & Jablonka, 2007, 2010, 2021). It is therefore likely that AL evolved 520 to 541 million years ago, presumably to guide taxis navigation, an even older adaptation defined as the navigational strategy to go *away from* (i.e., avoidance) or *toward* (i.e., approach) a valenced stimulus (see Bennett, 2021, for a discussion).

Several theories postulate that phylogenetically old adaptations (such as AL) have been conserved over evolutionary time and keep playing a significant role in modern human cognition. For instance, most evolutionarily-informed theories about neural architecture share the following notion: the more ancient an adaptation is, the more likely it is to have been recycled over evolutionary time – along with its neural substrates – as the building blocks of more recent cognitive computations (Anderson, 2014, 2016; Badcock et al., 2019; Dehaene, 2005; Dehaene & Cohen, 2007; Elimari & Lafargue, 2020). In a similar vein, the conservation of evolutionary old computations that bias rational reasoning, decision-making, and behaviors, is the core assumption of dual-process theories of human cognition (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2002, 2003; Gawronski & Creighton, 2013), a *meta*-theoretical approach that offers a figurative view of the mind as an interplay between evolutionary old, intuitive, unconscious, effortless, low-level processes (i.e., type I processes) to the most evolutionary recent, analytic, conscious, effortful, high-level processes (i.e., type II processes). Consistent with these theories, AL is implicated in several psychological phenomena such as (without being exhaustive) intergroup relationships and empathy (e.g., Cikara & Van Bavel, 2014; Melloni et al., 2014), reciprocal altruism (e.g., Rilling et al., 2002), moral cognition (e.g., FeldmanHall & Dunsmoor, 2019), sense of agency (e.g., Moore et al., 2011), selective social learning (e.g., Heyes, 2017), language acquisition (e.g., Ellis, 2006, 2008; Kachergis, 2012), synesthesia (e.g., Yon & Press, 2014), self-perception (e.g., Van Bavel & Cunningham, 2010), human and animal superstition (e.g., Beck & Forstmeier, 2007; Daprati et al., 2019; Skinner, 1948), or affective learning (Gawronski & Bodenhausen, 2006; Jones et al., 2009; Rydell & McConnell, 2006). This lends credence to the notion that AL is an evolutionary old, repeatedly repurposed adaptation. Our particular interest lies in the theorized role that AL plays in attitude formation (Gawronski & Bodenhausen, 2006; Jones et al., 2009; Rydell & McConnell, 2006), more specifically in its potential to shape attitudes outside of one's awareness in an automatic manner, through cognitive computations that operate independently of rational reasoning.

1.1. Propositional and dual-process theories of attitude formation

Research on attitude formation customarily involves *evaluative conditioning* paradigms, which capture the change in evaluation of a conditioned stimulus (CS) as a result of its repeated co-occurrence with an unconditioned stimulus (US) (Martin & Levey, 1978). The "*How*" question of evaluative conditioning however remains controversial (Houwer et al., 2005; Jones et al., 2010; Hofmann et al., 2010). While some authors argue that evaluative conditioning results from the non-automatic formation of explicit propositions (i.e., a mental statement that captures the specific manner in which two elements are linked, but also the degree of accuracy of that statement) about CS-US relations by a single, domain-general, awareness-dependent, propositional system (De Houwer, 2007, 2009; Mitchell et al., 2009), others have proposed that both propositional and AL processes are involved in attitude formation (Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006). For instance, the Associative Propositional Evaluation (APE) model (Gawronski & Bodenhausen, 2006) posits the existence of two separate systems that underpin attitude formation. The first is the AL system: a cluster of automatic, evolutionary old, low-level processes sensitive to mere spatiotemporal contiguity that transfer the affective charge of any US to a co-occurring CS, regardless of relational

information. The second is a propositional system characterized by controlled, evolutionary recent, high-level processes that produces explicit attitudes in the form of propositions crystallizing the relational information between stimuli.

Since AL processes do not code for complex relations (e.g., "*A starts/stops B*", "*A triggers/prevents B*", "*A is the same/opposite of B*"), the APE model predicts that AL produces implicit attitudes entirely mediated by automatic activation patterns that cannot be completely overcome by conscious, rational thinking. For instance, "*policemen*" leads to the automatic activation of "*crime*" though policemen *fight* crime, "*garbage collectors*" leads to the automatic activation of "*waste*" though garbage collectors *dispose of* waste, or "*doctors*" leads to the automatic activation of "*disease*", though doctors *cure* diseases. Consequently, the APE model predicts the possibility for the AL and propositional systems to give rise to two distinct sets of implicit and explicit attitudes, respectively. Given the blindness of the AL system to relational information and its strict focus on spatiotemporal contiguity, the APE model predicts that repeated co-occurrence of two stimuli linked by an antagonistic relation (e.g., "*A prevents B*", "*A stops B*") should prompt the two systems to produce two separate, incongruent, and mutually exclusive attitudes. Using this kind of incongruent evaluative conditioning, a series of experiments (e.g., Moran & Bar-Anan, 2013, 2020; Moran et al., 2015, 2016; Peters & Gawronski, 2011) have recently tested this prediction and confirmed that AL do automatically generate attitudes that can mismatch propositional attitudes. Moran and Bar-Anan (2013) have for instance compared congruently and incongruently conditioned attitudes toward target CSs (e.g., alien creatures that differed in color and head shape) that were presented with either positive (i.e., relaxing musical melody) or negative (i.e., horrifying human screams) USs. Participants were instructed to learn the specific relations between CSs and USs: in the congruent condition, CSs started the appearance of USs, whereas CSs stopped the occurrence of USs in the incongruent condition. The authors showed that explicit and implicit measures captured two seemingly independent attitudinal end-products: while explicit attitudes were consistent with relational information (e.g., preferred CSs that ended horrifying screams than CSs that ended beautiful music), implicit measures revealed a pervasive effect of mere co-occurrence (i.e., participants displayed an implicit preference for CSs paired with positive USs over CSs paired with negative USs, regardless of their relation).

Currently available evidence supporting the hypothesis of a parallel AL system unconsciously influencing attitude formation however remains conflicting (for a *meta*-analysis, see Hofmann et al., 2010). For instance, while Moran & Bar-Anan (2013) observed no effect of propositional processing on implicit evaluation, other studies have found that relational information could reverse (e.g., Gawronski et al., 2005) or attenuate (Zanon et al., 2012) implicit evaluations. Since behavioral research relying on explicit and/or implicit measures tend to provide mixed evidence, we propose a third path with the investigation of neural correlates of attitudes. To the best of our knowledge, there has been no attempt to advance the debate using neuroscientific methods.

1.2. A neural model of cognitive control

The two systems of the APE model are separate yet not independent: a collection of top-down and bottom-up processes ensure the mutual regulation and communication between the two systems (Gawronski & Bodenhausen, 2006). Several other theories posit the existence of a cognitive control system: a suite of supervisory or executive mechanisms subserving the continuous evaluation of competing representations or outcomes, the on-line maintenance of relevant information, and the top-down regulation of prepotent responses (Kahneman, 2003; Lieberman et al., 2002; Miller & Cohen, 2001; Norman & Shallice, 1986; Shiffrin & Schneider, 1977; Umilta, 1988). Starting from the early 90′ s (e.g., Dehaene et al., 1994; Falkenstein et al., 1991; Gehring et al., 1993), numerous studies have relied on EEG to investigate the neural correlates of key functions of cognitive control such as performance monitoring, conflict management, and error detection. The main finding was the existence of an event-related potential dubbed *error-related negativity* (ERN, Gehring et al., 1993), a negative deflection that follows error commission in a vast array of psychological tasks.

Descriptively, the ERN occurs at about the same time as motor response (or slightly earlier) during incorrect trials, peaks between 0 ms and 100 ms after motor response, and is quickly followed by a large positive wave labeled *error positivity* (Pe), which peaks between 200 ms and 400 ms. The ERN has a frontocentral distribution and is detected along the midline at electrode sites Fz, FCz, or Cz, while the Pe has a slightly more posterior topography (electrode sites CPz or Pz). Given its latency (i.e., starting around 0 ms), the ERN is described as an index of automatic error processing, while the Pe reflects the conscious appraisal of error processing (Endrass et al., 2005; Nieuwenhuis et al., 2001). Interestingly, correct trials also elicit a similar yet much smaller response-locked frontocentral negative wave dubbed the correct-response negativity (CRN). Consequently, several authors have relied on the so-called ΔERN (i.e., the difference between the ERN and the CRN) as a way to isolate error-specific correlates from generic responses neural activity. The amplitude of the ERN varies along with experimental conditions pertaining to error processing. For instance, ERN amplitudes are greater when errors are more costly (Hajcak et al., 2005; Holroyd et al., 2004), when focus is made on accuracy rather than speed (Gehring et al., 1993), when the perceived certainty that an error has been committed is higher, regardless of actual accuracy (Scheffers & Coles, 2000), when individuals know their performance is being subject to scrutiny (Hajcak et al, 2005; Meyer et al., 2019), after participants had restored their depleted cognitive resources by spending time in nature (LoTemplio et al., 2020), or when conflict between target stimuli and distractors is lower (Danielmeier et al., 2009).

1.3. The functional significance of error-related activity

Several theories proposed an explanation for the brain's ability to perform such an automatic and early error detection, most of which rest on the fundamental premise that the brain hosts a comparator system endowed with a weighting function that indexes degrees of divergence between expected and actual outcomes. For instance, the mismatch theory (Coles et al., 2001; Gehring et al., 1993) suggests that the ERN reflects the detection of discrepancies between intended and effective responses by the anterior cingulate cortex. In time-sensitive circumstances (as it is for instance the case in reaction time tasks), actions are sometimes initiated before all the information necessary to pinpoint the correct answer are gathered, thus

leading to a premature erroneous response. During such error commission, an "efference copy" is created and subsequently communicated to the comparator system that checks whether the efference copy matches the representation of a correct response derived from the further, continued, processing of the presented stimulus (Coles et al., 2001). In summary, the mismatch theory regards the ERN as an index of discrepancy between an overhasty erroneous action and the overdue representation of a correct response. In a similar vein, the reinforcement learning theory (Holroyd & Coles, 2002) conceives the ERN as a neural marker of expectancy violation: when the outcomes of an action are worse than expected, the midbrain dopamine system carries a signal to the anterior cingulate cortex to facilitate more adaptive motor programs. More recently, Alexander and Brown (2010) expanded on this expectancy violation framework with a prediction of response outcome theory (Alexander & Brown, 2010). This theory proposes that a neural system is implicated in the prediction of outcomes associated with planned actions, and the subsequent monitoring of discrepancies between predicted and actual outcomes. It is worth noting that, while consciousness of what constitutes the appropriate action and its consequences is necessary for the ERN to emerge, consciousness of the actual motor response is dispensable (Dehaene, 2018). Nieuwenhuis and colleagues (2001) have for instance shown using an antisaccade task (during which participants tend to produce several erroneous reflexive saccades that remain subjectively unnoticed) that incorrect saccades were systematically followed by an ERN regardless of error awareness, while the Pe amplitude positively correlated with the error awareness reported by the participants. Endrass and colleagues (2005) have observed the same impact of error awareness on Pe but not ERN amplitude with a saccade countermanding task. In other words, unlike post-error negativity, post-error positivity depends not on the valid representations of correct responses or expected outcomes, but on the level of awareness that an error has been committed.

1.4. The present research

The current study expands on the debate between single and dual-process theories of attitude formation by exploring the neural correlates associated with top-down control of CSs. More specifically, we investigate how distinct conditioning procedures (i.e., AL and propositional computations construing harmonious *vs* antagonistic attitudes) lead to differentials in error-related neural activity. Given the old age of valence, taxis navigation, and AL, as well as the overrepresentation of AL processes in several cognitive domains, we err on the side of dual-process theories.

We designed a rather straightforward demonstration: if evolutionary old, AL computations do produce outside of one's awareness implicit attitudes that are paralleled with those generated by the conscious propositional system, then there is likely to be a significant difference in the error-monitoring processes following responses toward congruent and incongruent stimuli (beyond the well-replicated effects of incongruence on subjective valence and behavioral performance during time-reaction tasks, e.g., Moran & Bar- Anan, 2013, 2020; Moran et al., 2015, 2016; Peters & Gawronski, 2011). More specifically, we hypothesize that ERN amplitudes will vary as a function of stimulus congruency, with diminished amplitudes for incongruent stimuli. Indeed, attitudes are fundamentally guides to action: they help navigating the world by facilitating adapted motor programs towards valenced stimuli (Allport, 1935;

Carruthers, 2017; Chaiklin, 2011; Jain, 2014; Olson & Fazio, 2008; Petty et al., 2007; Van Overwalle & Schibler, 2005; Shrigley, 1990). Therefore, when one produces motor responses involving CSs, any conflict in attitudes would result in impaired information-processing downstream of attitudinal cognition. Following the mismatch, reinforcement-learning, and prediction of response outcomes theories, we predict that a conflict of attitudes should decrease the ability of the comparator system to construe (1) a representation of the correct response and (2) a coherent set of predictions about post-response outcomes. This would result in the subsequent impaired ability of the comparator system to properly weigh expected outcomes against actual outcomes.

Though our hypotheses mainly concern ERN, we will also investigate the effect of congruence on Pe amplitudes in an exploratory fashion. Since error-positivity is viewed as an index of error awareness, we propose the rather intuitive hypothesis that attitudinal conflict will prevent performance monitoring processes to fully gain conscious access to error commission. We expect so for the same reason we evoked before: if dual-process theories are accurate in their claim that the brain hosts two attitudinal minds, and if only one of them believes to be wrong, only a fraction of the "alarm" error-signal should be triggered. Conscious access in the brain can be understood in the form of accumulation of information-related signal above consciousness threshold (Del cul et al., 2007; 2009). If such is the case, attitudinal conflict should disrupt accumulation of error-related information signal and jeopardize conscious access to error. Thus, we expect Pe amplitudes to correlate negatively with incongruence.

2. METHODS

2.1. Participants

Sample size calculation was computed using G*Power 3.1.9.7 (Faul et al., 2009) for an effect size of 0.54 (smallest effect size reported for incongruence-based differences in ERN amplitudes, Danielmeier et al., 2009), statistical power of 0.80, and a one-tailed hypothesis (congruent > incongruent). Minimal sample size was determined at 23 participants. To anticipate loss of data due to technical difficulties and insufficient amounts of error trials, we chose to raise this number to 40. Forty French-speaking participants (24 females, 16 males) between the age of 18 and 32 (M = 23.65, SD = 3.91) volunteered to the experiment. Participants were recruited from the general population with the use of flyers distributed on university campus. Participants had normal or corrected-to-normal vision. The study was designed in accordance with the Declaration of Helsinki and all participants gave their written informed consent after receiving a full description of the study. All participants were informed they could withdraw from the experiment at any time.

2.2. Procedure

Participants were comfortably seated in an armchair placed in a dim lit and sound attenuated room in front of a 17″ screen computer. The experiment was implemented using E-Prime 2 Professional (Psychology Software Tools, Pittsburgh, PA, USA) and consisted of two phases: (1) a conditioning phase where attitude formation was induced, and (2) a data-recording phase where CSs were maneuvered and evaluated in various ways by the participants. Three types of

data were extracted from the second phase: subjective data, behavioral data, and electrophysiological data. Schematic of the experimental design is provided in Figure 1.

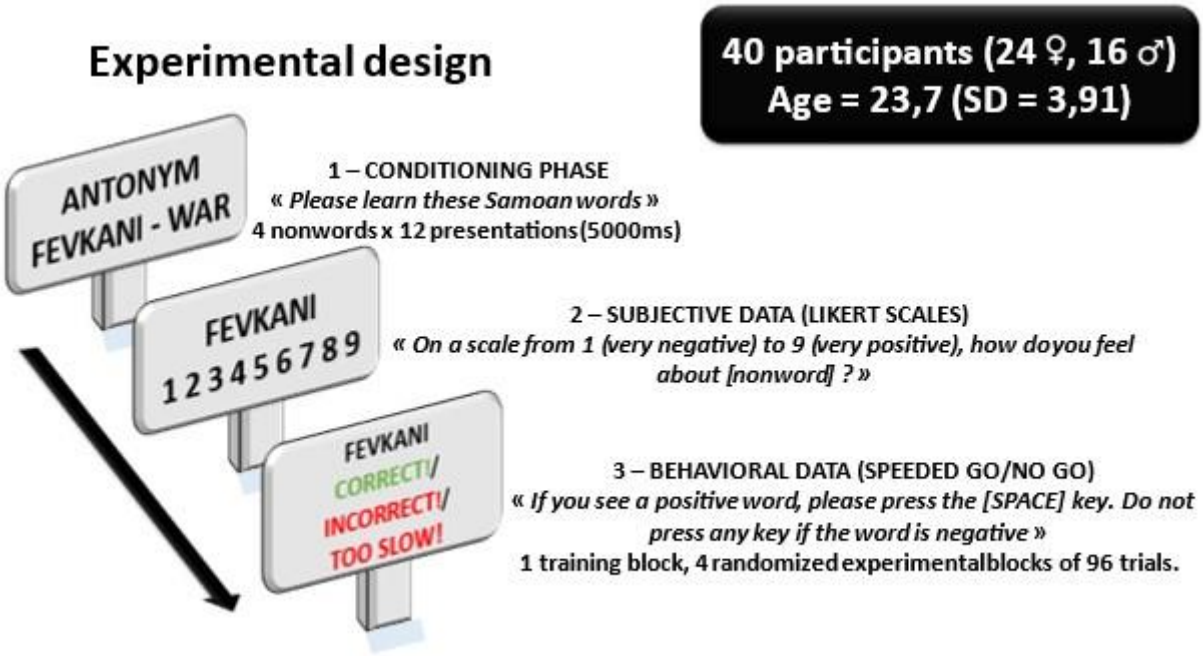

Figure 1. *Graphical representation of the experimental design. The conditioning phase consisted of 12 presentations of 4 nonwords that participants were instructed to learn. The second phase consisted of subjective evaluation reports (Likert scales) and a Speeded Go/No Go. EEG data were collected during the Speeded Go/No Go only.*

The conditioning phase was submitted in the form of a learning task inspired from Zanon and colleagues (2014), in which participants had to memorize the definition of four seven-letters nonwords (i.e., LOKANTA, FEVKANI, POIMATA, and NITAIKI) presented as words from the Samoan language. Nonwords were either learned with their direct translation or with the translation of their antonym, in which case the participant had to infer the actual meaning (e.g., FEVKANI = WAR → PEACE). This was illustrated with the following example: "*if the Samoan word ROGAIDI is presented with the French word "left" in the antonym condition, the real meaning is "right". The same would apply for "up" → "down" or "slow" → "fast" for example"*. The four CSs (i.e., nonwords) and four USs (i.e., French words, *HEUREUX* or *HAPPY*, *ATROCITʹE* or *ATROCITY*, *BONHEUR* or *BLISS*, & *GUERRE* or *WAR*) were assigned to each other in a pseudo-random fashion. The CS/US pairs were presented 12 times each for 5000 ms in a random order. The CSs were considered congruent when their valence was the same as the US valence (i.e., direct translation), and incongruent when the CSs valence were the opposite of USs valence (e.g., "*fevkani - war*" in the antonym condition, where participants learned that "*fevkani*" translated into "*peace*" but were presented with repeated "*fevkani - war*" pairings). Participants then completed a short multiple-choice questionnaire to assess whether they managed to memorize each nonword. No errors were committed by any of the participants. Moreover, no participant reported any difficulty memorizing the four nonwords, nor any significant additional cognitive load associated with deducing the actual meaning of nonwords.

2.3. Data collection

2.3.1. Subjective data

Subjective data was recorded using a series of Likert scales. Participants rated each nonword by answering the following question *"On a scale from 1 (very negative) to 9 (very positive), how do you feel about the word [CS]?*". Since the Go/No-Go is intrinsically a categorization task that can potentially induce a counterconditioning of the CSs, explicit valence ratings were systematically presented before the inhibition task (Figure 1).

2.3.2. Behavioral data

Behavioral data was recorded using a modified Speeded Go/No-Go. The task started with a first instruction screen reading "*You are about to start a categorization task during which you will have to either press the [SPACE] key or refrain from pressing it, depending on the positive or negative valence of a word: if the word displayed on the screen is positive (*or negative*), then press the [SPACE] key, but if the word is negative* (or positive), *do not press any key. Both the Samoan words you just learned and French words will be presented. A training session will help you get accustomed with the task. Please press any key to start the training session*". Participants then went through a training session consisting of 40 trials presented in a random order. A fixation cross appeared for 800 ms, followed by a semantic stimulus (maximal time = 3000 ms). Correct responses were followed by a feedback screen displayed for 1500 ms reading "*CORRECT!*" in a green font, while errors were followed with a feedback screen reading "*INCORRECT!*" in a red font. During the training session, the average response time of each participant was recorded and established as an individualized performance baseline that had to be surpassed during the experimental session.

The experimental session was similar to the training session, except that fixation cross and feedback duration dropped to 500 ms, and maximal presentation time was reduced to 2000 ms. On top on instructing participants to respond as quickly as possible, we implemented a rule-based feedback that suppressed normal feedback screen (i.e., "*CORRECT/INCORRECT!*" message, response time, accuracy) and replaced it with another screen reading "*TROP LENT!*" (or "*too slow!*") if response time was higher than 90% of the average latency recorded during the training phase. Therefore, the pace of the task was individually fixed so that participants had to respond 10% faster than their normal speed, thus maximizing the number of errors without overwhelmingly exceeding participants' capacities. EEG data were not recorded for slow trials so as to avoid conflating generic correct responses and correct responses followed by an intrinsically negative feedback).

Normal French words were presented along with the conditioned CSs to determine the baseline neural activity associated with processing standard positive and negative words. To keep the same number and probability of occurrence of words and nonwords, four standard words (i.e., two positive: "*JOYEUSE*" or "*happy*", and "*DOUCEUR*" or "*gentleness*", and two negative: "*MUTILER*" or "*mutilate*" and "*INFECT´E*" or "*infected*") were selected according to their valence. The experimental session comprised 4 blocks (2 blocks requiring categorization of positive stimuli, 2 others of negative stimuli) of randomly ordered 96 trials, for a total of 384 trials. We purposedly avoided increasing the total number of trials over 384 to prevent any potential implicit counterconditioning arising from the repeated categorization of nonwords as

either positive or negative. Trials requiring motor responses were presented 66% of the time. Blocks were separated by pause screens without pre-programmed duration and participants were instructed to rest as long as they wished. Rather than giving instructions to avoid blinking, participants were advised to rest their eyes during pauses to minimize the average number of blinks per minute. Average response time for correct hits and errors, as well as average number of errors were computed for each condition (i.e., Congruent, Incongruent, Standard words).

2.3.3. Electrophysiological data

Electrophysiological data were collected at 1000 Hz using Brainvision PyCorder (Brain Products GmbH, Gilching, Germany) and a 32-channel system (Brainvision actiCHamp, Brain Products GmbH. Gilching, Germany) with Ag/AgCl active electrodes positioned on a cap (*actiCAP*, EASYCAP, GmbH) at positions Fz, F3, F7, FT9, FC5, FC1, C3, T7, TP9, CP5, CP1, Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, CP6, CP2, Cz, C4, T8, FT10, FC6, FC2, F4, and F8. Electrodes Fp1 and Fp2 were derived to record vertical and horizontal electrooculography, respectively. Signal was referenced to a ground electrode placed at position Fpz. A preprocessing procedure was performed offline using EEGLAB (Delorme & Makeig, 2004) and Brainstorm (Tadel et al., 2011) and comprised (1) downsampling to 250 Hz, (2) basic FIR filter between 0.1 Hz and 35 Hz, (3) visual inspection and subsequent removal of noisy segments/bad channels, (4) independent component analysis followed with semi-automatic detection and correction of artefacts using the ADJUST algorithm (Mognon, Jovicich, Bruzzone, & Buiatti, 2011), (5) re-referencing to average, and (6) creation of EEG segments from 150 ms to 750 ms after motor response onset (including baseline correction of the mean activity 150 ms to 50 ms before response onset). Correct and error trials were averaged separately using mean amplitude between 50 ms and 100 ms. To quantify error-specific neural activity, ΔERN and ΔPe scores were calculated as the difference between average correct-response (CRN) and error-related (ERN) activities. The ΔERN was quantified as the mean amplitude between 50 ms and 100 ms at Cz, where error-related activity was maximal, while ΔPe was quantified as the average activity between 200 ms and 500 ms at Pz.

2.4. Data analyses

2.4.1. Subjective data

Repeated-measures analyses of variance (ANOVA) were used to analyze differences between explicit subjective ratings with Valence (positive, negative) and Congruence (congruent, incongruent) as within-subject factors. We hypothesize a main effect of Valence, attesting to the efficacy of the conditioning procedure, and a Valence × Congruence interaction effect reflecting a variation of the effect of valence as a function of congruence, with incongruent words being evaluated as more neutral than their congruent counterparts.

2.4.2. Behavioral data

One-way analyses of variance (ANOVA) were used to investigate differences between congruent CSs, incongruent CSs and standard words with regards to average response times during correct and error trials, as well as average number of errors. Paired t-tests were used as planned contrasts (Hager, 2002) to further explore *a priori* predictions about the effect of

condition on behavioral data. We hypothesize that incongruent CSs will be associated with significantly longer RTs (for correct and error trials) and error rates than both congruent CSs and standard words. Similarly, since congruent CSs were only recently learned, they were presumably harder to mentally manipulate than long-known words. For this reason, we also hypothesize that congruent CSs will be associated with longer RTs and error rates than standard words. As our hypotheses are directed, we chose to set a significance level at $\alpha = 0.1$.

2.4.3. Electrophysiological data

Of the initial 40 participants, 3 were excluded from the sample for corrupted data and 32 made enough errors in both the congruent and incongruent conditions to allow for statistical analyses. Of these 32 participants, 26 (15 females, 11 males, age = 23.92, SD = 3.78) also made enough errors during standard words categorization. A one-way ANOVA was therefore computed on these 26 participants to explore the variations of amplitudes of ERP components (i.e., ΔERN, ΔPe) as a function of Condition (congruent, incongruent, standard words). Following Thigpen and colleagues (2017), internal consistencies of both components were assessed by calculating Cronbach's alphas with condition-averaged ERPs (k = 3 conditions). Both ΔERN ($\alpha = 0.739$) and ΔPe ($\alpha = 0.839$) presented with acceptable reliability. We specifically relied on paired t-tests as planned contrasts (Hager, 2002) as a way to analyze *a priori* predictions about the effect of each condition on ERP amplitudes, all the while adjusting for the variability in the minimal number of usable erroneous trials. We hypothesize that incongruent CSs will be associated with diminished error-related amplitudes compared to both congruent CSs and standard words. Once again, since congruent CSs were recently learned, it is logical to assume for a lesser ability of the comparator system to construe the mental representations of both the correct response and expected outcomes. We thus hypothesize a significant decrease in ΔERN and ΔPe for congruent CSs when compared to standard words. Given our directed hypotheses, we set a significance level at $\alpha = 0.1$.

3. RESULTS

3.1. Effect of valence and congruence on subjective data

Mean explicit evaluation scores are illustrated in Figure 2. A first 2 × 2 Analysis of Variance (ANOVA) was performed with Congruence (congruent, incongruent) and Valence (positive, negative) as within-subject factors to determine the effect of the conditioning procedure on explicit evaluation. The ANOVA found no significant main effect of Congruence [$F(1, 39) = 0.23$, $p = .77$, $\eta 2 = 0.0002$], and a significant main effect of Valence [$F(1, 39) = 62.4$, $p < .001$, $\eta 2 = 0.62$], reflecting a more positive attitude toward CSs that had a positive rather than a negative meaning, thus confirming the efficacy of the conditioning procedure, see Figure 2.

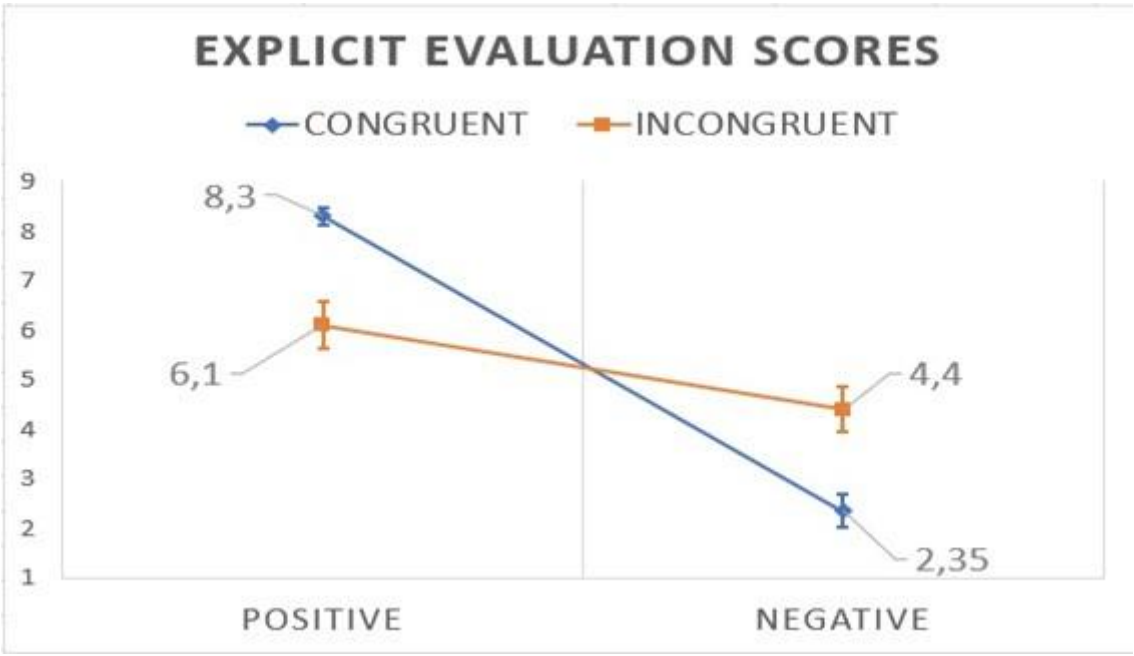

Figure 2. *Mean explicit ratings for CSs as a function of valence and congruence on a Likert scale of affective rating from 1 (very negative) to 9 (very positive). Data show that positive congruent CSs were perceived as more positive than their incongruent counterpart. On the other hand, congruence resulted for negative stimuli in a more negative subjective evaluation of congruent CSs.*

A significant Congruence × Valence interaction effect was found [$F(1, 39) = 29.55$, $p < .001$, $\eta 2 = 0.43$], reflecting a reversal effect of congruence on the outcome of the conditioning procedure as a function of valence: positive congruent CSs were rated more positively than positive incongruent CSs while negative congruent CSs were rated more negatively than negative incongruent CSs. Paired *t*-tests confirmed that pattern of results, as congruent CSs (mean valence = 8.3) were perceived as significantly more positive than incongruent CSs (valence = 6.1) for positive stimuli [$t(39) = 4.56$, $p < .001$, $d = 0.99$], while congruent CSs (valence = 2.35) were evaluated more negatively than incongruent CSs (valence = 4.4) for negative stimuli [$t(39) = 4.51$, $p < .001$, $d = 0.81$].

3.2. Effect of congruence on behavioral data

Differences in RT (for both correct and error trials) and error rates were assessed with three separate One-way ANOVAs with Condition (Congruent, Incongruent, Standard words) as independent variable. Average RTs during correct and error trials for all conditions are reported in Figure 3. The first ANOVA revealed a significant difference between conditions in terms of RT [$F(2, 117) = 21.56$, $p < .001$, $\eta 2 = 0.27$]. Paired t-tests corroborated this finding, with incongruent CSs (518.05 ms) being categorized slower than both congruent CSs (489.01 ms) [$t(39) = 4.6$, $p < .001$, d = 1.39] and standard words (470.43 ms) [$t(39) = 8.76$, $p < .001$, $d = 0.73$]; and congruent CSs being categorized slower than standard words [$t(39) = 3.68$, $p < .001$, $d = 0.58$], see Figure 3.

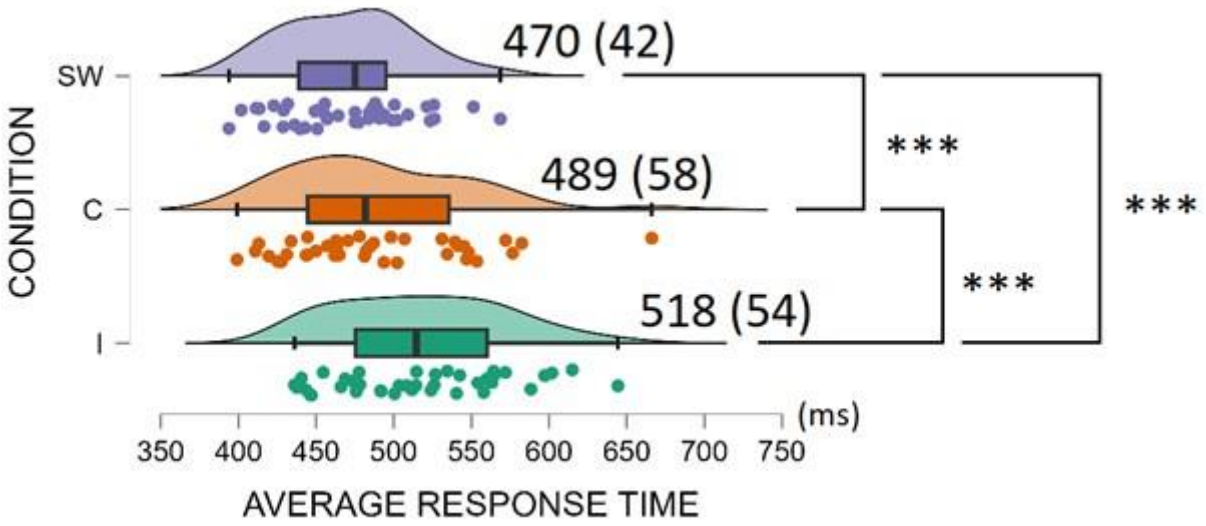

Figure 3. *Means (standard deviation) and differences between conditions in average response times. Incongruent CSs (I) took significantly more time to be processed than congruent CSs (C) and standard words (SW). Congruent CSs were themselves associated with increased RTs when compared with standard words.*

A similar pattern of results was observed for RTs during error commission. A second One-way ANOVA revealed that response time during error trials also varied as a function of Condition [$F(2, 115) = 15.22$, $p < .001$, $\eta 2 = 0.21$]. Paired t-tests confirmed that stimuli all differed from each other: error RTs were longer for incongruent CSs (496.66 ms) than for congruent CSs (477.56 ms) [$t(38) = 1.95$, $p = .059$, $d = 0.31$] and standard words (417.38 ms) [$t(38) = 7.75$, $p < .001$, $d = 1.24$], while error RTs were longer for congruent CSs than standard words [$t(37) = 7.22$, $p < .001$, $d = 1.17$].

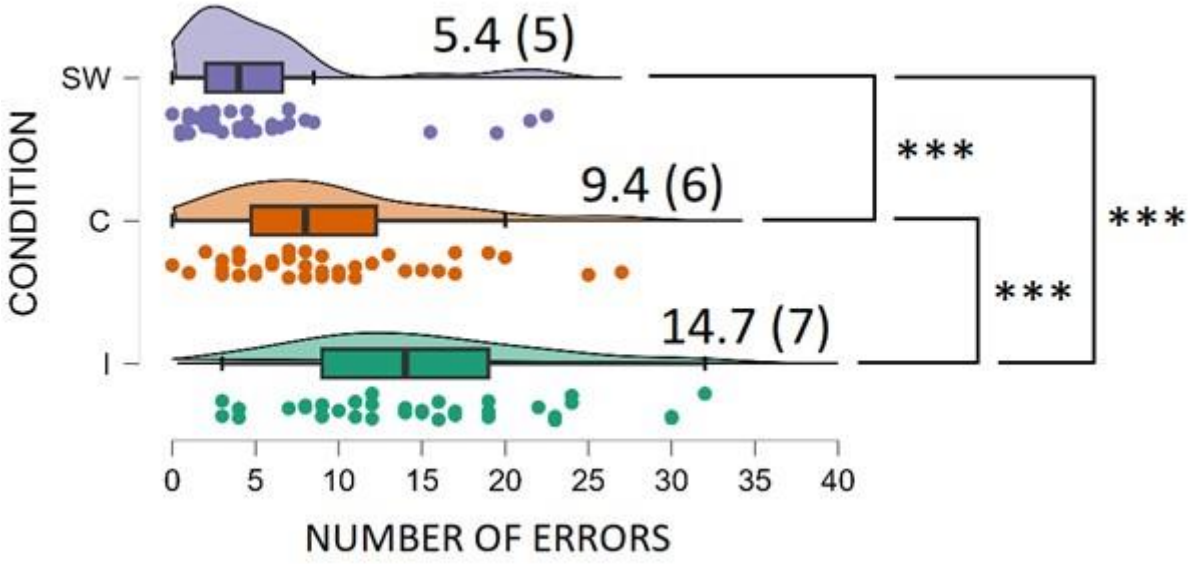

Figure 4. *Means (standard deviation) and differences between conditions in number of errors. Incongruent CSs (I) elicited significantly more errors than congruent CSs (C). Congruent CSs elicited more errors than standard words (SW). It is worth noting that, in order to present the same number of nonwords and standard words, four of the latter category were presented during the task: The average*

number of errors reported above were divided by two to account for that fact, however, twice as many error trials were available when processing EEG data.

The average number of errors made per condition is reported in Figure 4. On average, incongruent CSs elicited 14.7 errors, congruent CSs elicited 9.4 errors, and standard words elicited 5.4 errors (please note that the reported number of errors is here divided by two to account for probability of occurrence of standard words during the task, and that twice as many trials were available during EEG data processing). The second ANOVA showed that conditions also affected the number of errors made [$F(2, 117) = 8.69$, $p < .001$, $\eta 2 = 0.13$]. Alpha-corrected paired *t*-tests confirmed that participants categorized incongruent CSs less accurately than both congruent CSs [$t(39) = 7.38$, $p < .001$, $d = 0.1.17$] and standard words [$t(39) = 11.39$, $p < .001$, $d = 1.8$]. Once again, behavioral performance was better for standard words than congruent CSs, with higher accuracy during standard words categorization [$t(39) = 5.36$, $p < .001$, $d = 0.85$].

3.3. Effect of congruence on electrophysiological data

Grand average waveforms of ΔERN (electrode Cz) and ΔPe (electrode Pz) are presented in Figures 5 and 6, respectively. A one-way ANOVA was computed to explore differences in automatic error-detection between conditions. The ΔERN amplitudes showed a significant effect of Condition [$F(2, 87) = 5.27$, $p = .007$], with a mean amplitude of 4.56 μV (SD = 3.45) for incongruent CSs, 7.41 μV (SD = 6.67) for congruent CSs, and 10.15 μV (SD = 8.88) for standard words. Confirming our main predictions, incongruent CSs were associated with decreased ΔERN amplitudes compared to both congruent CSs [$t(31) = 3.03$, $p = .005$, $d = 0.54$] and standard words [$t(25) = 4.19$, $p < .001$, $d = 0.82$]. When compared with standard words, congruent CSs were themselves associated with significantly smaller ΔERN amplitudes [$t(25) = 2.78$, $p = .01$, $d = 0.55$].

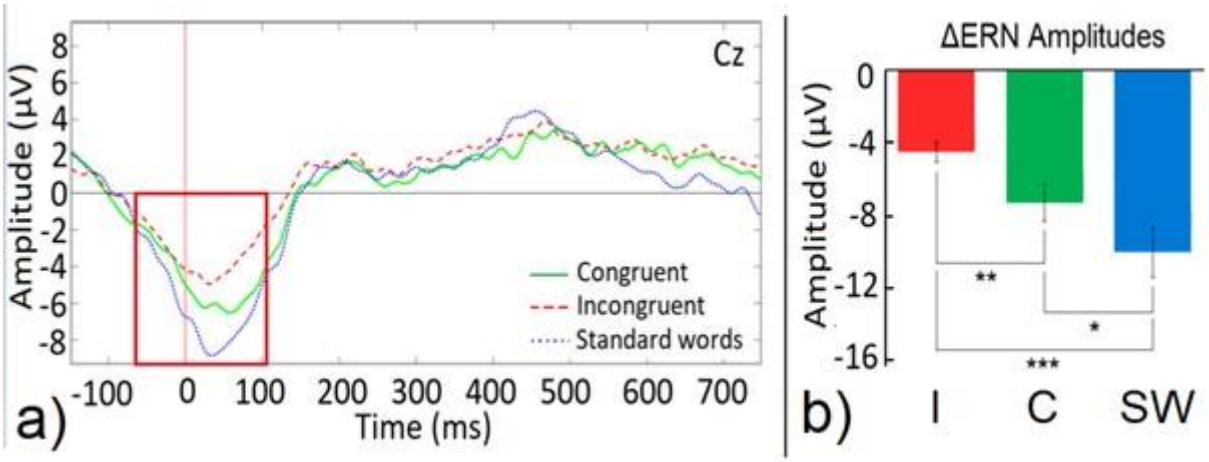

Figure 5. *(a) Grand average waveforms at electrode Cz. Continuous green lines picture congruent CSs, dashed red lines picture incongruent CSs, and dotted blue lines picture standard words. The red frame marks the time-window of interest. (b) Differences in ΔERN amplitudes between conditions. Incongruent CSs (I) were associated with significantly smaller ΔERN when compared with both congruent CSs (C) and standard words (SW). Congruent CSs were also associated with diminished ΔERN amplitudes compared to standard words. Note: *p < .05, **p < .01, ***p < .001.*

To assess differences in conscious awareness of error commission (i.e., ΔPe) between conditions, a final one-way ANOVA was computed. Consistent with our main predictions, Condition significantly affected ΔPe amplitudes [$F(2, 87) = 4.6$, $p = .013$, $\eta 2 = 0.1$], with decreased amplitudes for incongruent CSs (5.01 μV) compared to congruent CSs (8.12 μV) [$t(31) = 4.51$, $p < .001$, $d = 0.8$] and standard words (11.83 μV) [$t(25) = 4.4$, $p < .001$, $d = 0.87$]. Similarly, congruent CSs elicited significantly more error-positivity than standard words [$t(25) = 2.43$, $p = .023$, $d = 0.48$] (see Figure 6).

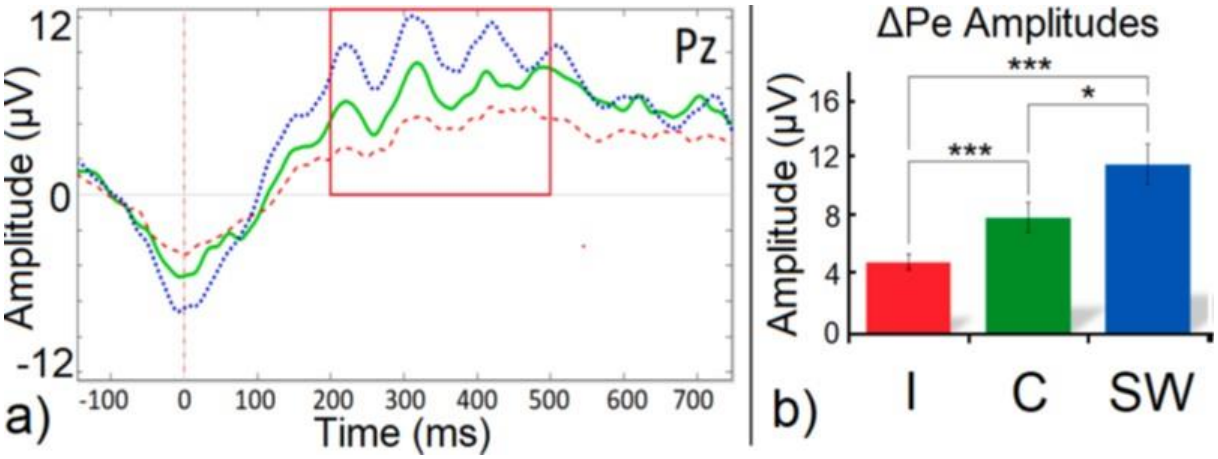

Figure 6. *(a) Grand average waveforms at electrode Pz. Continuous green lines picture congruent CSs, dashed red lines picture incongruent CSs, and dotted blue lines picture standard words. The red frame marks the time-window of interest. (b) Differences in ΔPe amplitudes between conditions. As it was observed for ΔERN amplitudes, incongruent CSs (I) were associated with significantly smaller ΔPe when compared to both standard words (SW) and congruent CSs (C); while congruent CSs were themselves associated with diminished ΔPe amplitudes compared to standard words. Note: *p < .05, **p < .01, ***p < .001.*

4. DISCUSSION

The present study sought to provide evidence for the activity of a vestigial associative-learning (AL) system during attitude formation in modern human cognition. Our experiment was constructed on the premise that it was possible, knowing the computational nature of both the AL and propositional systems, to condition mutually exclusive attitudes toward a same stimulus. We thus conditioned stimuli in such a way that evolutionary old AL computations and evolutionary recent propositional computations worked either in harmony (i.e., generating congruent attitudes) or in disharmony (i.e., generating incongruent attitudes), and investigated how explicit evaluation, inhibition performance, and neural correlates of error monitoring varied as a function of said harmony. Confirming previous work (e.g., Moran et al., 2016, 2017), subjective evaluations of CSs were affected by incongruence, as incongruent CSs were evaluated as more neutral than their congruent counterparts. This suggests that AL computations processing mere spatiotemporal contiguity between CSs and USs had an effect on explicit ratings above and beyond the actual meaning of the nonwords derived from relational information. In other words, implicit antagonistic attitudes generated by AL computations downregulated explicit evaluations and thus, perceived valence. In this study, we used a Speeded Go/No Go task as a way to assess inhibition performance of CSs, with a focus on response times (of both correct and error trials) and number of errors as proxy markers of

conflict between implicit and explicit attitudes (based on the same principle used for instance in the IAT, Greenwald et al., 2001). Standard words were categorized along with CSs to have a sense of information processing of classic everyday semantic stimuli. On average, incongruent CSs were associated with significantly longer response times and more errors than both congruent CSs and standard words, suggesting once again that AL computations processed information in an automatic and parallel fashion, and construed implicit attitudes (independent of intention and awareness) that affected explicit attitudes based on the propositional processing of complex relational information. The same conclusion can be drawn from electrophysiological data: incongruence was associated with decreased neural correlates of error-processing. Overall, the incongruent evaluative conditioning procedure did influence explicit ratings, inhibition performance, and neural correlates of both error monitoring and conscious awareness of error commission. Our results support the claim put forward by dual-process theories of attitude formation that a vestigial AL system sensitive to mere spatiotemporal contiguity is still actively implicated in the dynamic adaptation of attitudes toward environmental cues (Chaiken & Trope, 1999; Gawronski & Bodenhausen, 2006; Morewedge & Kahneman, 2009; Petty & Caccioppo, 1986; Rydell & McConnell, 2006). This system produces attitudes outside of one's awareness by automatically detecting and processing recurring patterns of association in the environment, irrespective of rule-based, conscious, rational thinking. To our knowledge, the results presented here constitute the first evidence based on electrophysiological data in favor of dual-process theories of attitude formation.

4.1. On the functional role of error-related negativity

Our results also reinforce the validity of "comparator" theories of the ERN (e.g., Mismatch theory, Coles et al., 2001; Reinforcement learning theory, Holroyd & Coles, 2002; Prediction of response outcome theory, Alexander & Brown, 2010). These theories rest of the fundamental assumption that the ERN results from the automatic detection of discrepancies between expected and actual outcomes. However, such discrepancy detection mechanically necessitates for the comparator system to access (1) a coherent representation of the expected correct response and (2) a balanced and uniform set of predictions about action outcomes. We hypothesized that the conflict of attitudes that resulted from our paradigm would undermine both mental constructs (i.e., correct response, outcome prediction), and our results indicated that attitudinal conflict indeed impaired the comparator system's ability to detect discrepancies, as reflected in the diminished ERN and Pe amplitudes observed in our study. It is also worth noting that our results are consistent with "affective" theories of the ERN (Luu et al., 2003; Tucker et al., 1999), which consider this neural marker as a index of aversiveness to errors. In a way, the fact that the motor program associated with errors (understood from the point of view of explicit attitudes of the propositional system) is actually compatible with implicit attitudes might have reduced the affective load carried by errors: part of the brain considered that the error was a correct response and that the motor program was thus appropriate, which reduced the overall aversiveness of errors in the incongruent condition. Our results are also compatible with conflict-monitoring theories of the ERN (Botvinick et al., 2001; Carter et al., 1998; Yeung et al., 2004), which propose that the ERN arises from the increased activity within a response system attempting to manage a conflict between simultaneous yet mutually incompatible

response options. Following this framework, previous work has observed similar pattern of results as those presented in this study. For instance, Danielmeier and colleagues (2009) predicted and confirmed that high-conflict conditions were associated with lower ERN amplitudes in high-conflict conditions. Finally, our data provide useful insight into the mechanics of error-positivity. Indeed, one could have argued that Pe, a well-known index of error-awareness (Endrass et al., 2005; Nieuwenhuis et al., 2001), would have correlated with propositional, explicit, conscious evaluations only. In other words, that Pe amplitudes would remain unaffected by associative, implicit, unconscious attitudes. However, attitudinal conflict disrupted conscious access to error commission, as reflected in the significantly decreased Pe amplitudes. We grounded our hypotheses in models proposing that conscious access to information is dependent on the accumulation of signal above a consciousness threshold (e.g., Del cul et al., 2007; 2009). We expected attitudinal conflict to diminish the overall, whole-brain accumulation of error signal above consciousness threshold because part of the brain still believed the error to be the correct answer (as motor response mismatched explicit attitude but matched its implicit counterpart). Future work should however explore our claim more thoroughly, with more accurate indexes of error-signal accumulation as a function of attitudinal conflict.

4.2. Limitations and future directions

This study needs to be replicated to confirm the pattern of results we observed, especially since our study is not without limits. First, we have been dealing with a relatively low number of errors. This limit is directly linked to the originality of our paradigm: manually conditioning specific attitudes – rather than relying on stimuli that carry their own valence such as angry faces or cockroaches – intrinsically implies the risk to induce counterconditioning in the testing phase (by repeatedly categorizing an initially negative stimulus as positive hundreds of times for instance). For this reason, we deliberately kept the number of trials as low as possible in the Go/No-Go task. This however resulted in an average number of error trials going from 9.4 to 14.7 per condition. Though grand averages confirmed statistical analyses, replication studies could gain from multiplying sessions of conditioning procedures, so as to reinforce conditioned attitudes enough to allow for an extended period of testing. Another limit is the use of semantic stimuli, which might not be conducive of everyday, organic attitude formation. Future studies should test the same procedure on different types of stimuli such as humans or commercial products. The AL system might present with domain-specific variations in its degree of influence depending on stimulus nature. It is for instance worth wondering whether the same procedure would affect social cues in the same manner. Indeed, social cognition has long been described as one of the most complex and recently evolved sets of cognitive mechanisms, as well as the potential reason for the evolution of unique human intelligence (e.g., Bailey & Geary, 2009; Dunbar, 1998; Gavrilets & Vose, 2006; Holloway, 1967; Humphrey, 1976). One might therefore expect social cognition to function on the basis of more phylogenetically recent, flexible, conscious computations (Elimari & Lafargue, 2020), which would entail a higher dominance of the propositional system during social judgments. Overall, this study provides a new way to look at dual-process theories, as well as new avenues for testing hypotheses derived from both dual-process and evolutionary frameworks. Neural correlates such as the ERN and

Pe have a well-established literature that has extensively studied their specifics and features: they are ideal candidates for quantifying relations between brain reactions and subjectively perceived values. The search for more objective measures of cognitive phenomena is a goal shared by most researchers in Psychology. The present study opens the door for the development of new ways to exploit neural correlates as proxy measures for the quantification of subjective evaluation.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abramson, C. I. (1994). *A primer of invertebrate learning: The behavioral perspective*. American Psychological Association.

Alexander, W. H., & Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in cognitive science, 2*(4), 658–677.

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Anderson, M. L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, M. L. (2016). Pr´ecis of after phrenology: Neural reuse and the interactive brain. *Behavioral and Brain Sciences, 39*.

Ardiel, E. L., & Rankin, C. H. (2010). An elegant mind: Learning and memory in Caenorhabditis elegans. *Learning & memory, 17*(4), 191–201.

Badcock, P. B., Friston, K. J., Ramstead, M. J., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cogn. Affect. Behav. Neurosci., 19*, 1319–1351. https://doi.org/10.3758/s13415-019-00721-3

Bailey, D. H., & Geary, D. C. (2009). Hominid brain evolution. *Human Nature, 20*(1), 67–79.

Beck, J., & Forstmeier, W. (2007). Superstition and belief as inevitable by-products of an adaptive learning strategy. *Human Nature, 18*(1), 35–46.

Bennett, M. S. (2021). What behavioral abilities emerged at key milestones in human brain evolution? 13 hypotheses on the 600-million-year phylogenetic history of human intelligence. *Frontiers in Psychology, 12*.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review, 108*(3), 624.

Carter, C. S., Botvinick, M. M., & Cohen, J. D. (1999). The contribution of the anterior cingulate cortex to executive processes in cognition. *Reviews in the Neurosciences, 10*(1), 49–58.

Chaiklin, H. (2011). Attitudes, behavior, and social practice. *J. Soc. & Soc. Welfare, 38*, 31.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9*(3), 245–274.

Coles, M. G., Scheffers, M. K., & Holroyd, C. B. (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological psychology, 56*(3), 173–189.

Danielmeier, C., Wessel, J. R., Steinhauser, M., & Ullsperger, M. (2009). Modulation of the error-related negativity by response conflict. *Psychophysiology, 46*(6), 1288–1298.

Daprati, E., Sirigu, A., Desmurget, M., & Nico, D. (2019). Superstitious beliefs and the associative mind. *Consciousness and Cognition, 75*, Article 102822.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish journal of psychology, 10*(2), 230–241.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*(1), 1–20.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. *From Monkey Brain to Human Brain*, 133–157.

Dehaene, S. (2018). The error-related negativity, self-monitoring, and consciousness. *Perspectives on Psychological Science, 13*(2), 161–165.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron, 56*(2), 384–398.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science, 5*(5), 303–305.

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS biology, 5*(10), e260.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain, 132*(9), 2531–2540.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods, 134*(1), 9–21.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews, 6*(5), 178–190.

Elimari, N., & Lafargue, G. (2020). Network neuroscience and the adapted mind: Rethinking the role of network theories in evolutionary psychology. *Frontiers in psychology, 11*, 2546.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics, 27*(1), 1–24.

Ellis, N. C. (2008). *Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate*. Routledge/ Taylor & Francis Group.

Endrass, T., Franke, C., & Kathmann, N. (2005). Error awareness in a saccade countermanding task. *Journal of Psychophysiology, 19*(4), 275–280.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223–241.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology, 78*(6), 447–455.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

FeldmanHall, O., & Dunsmoor, J. E. (2019). Viewing adaptive social choice through the lens of associative learning. *Perspectives on Psychological Science, 14*(2), 175–196.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: Their relative effectiveness as a function of the reinforcer. *Science, 160*(3829), 794–795.

Gavrilets, S., & Vose, A. (2006). The dynamics of Machiavellian intelligence. *Proceedings of the National Academy of Sciences, 103*(45), 16823–16828.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological bulletin, 132*(5), 692.

Gawronski, B., & Creighton, L. A. (2013). Dual process theories. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition*. New York: Oxford University Press.

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology, 41*(6), 618–626.

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological science, 4*(6), 385–390.

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes, 25*(3), 363.

Ginsburg, S., & Jablonka, E. (2007). The transition to experiencing: I. Limited learning and limited experiencing. *Biological Theory, 2*(3), 218–230.

Ginsburg, S., & Jablonka, E. (2010). The evolution of associative learning: A factor in the Cambrian explosion. *Journal of theoretical biology, 266*(1), 11–20.

Ginsburg, S., & Jablonka, E. (2021). Evolutionary transitions in learning and cognition. *Philosophical Transactions of the Royal Society B, 376*(1821), 20190766.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

Hager, W. (2002). The examination of psychological hypotheses by planned contrasts referring to two-factor interactions in fixed-effects ANOVA. *Method Psychol Res Online, 7*, 49–77.

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology, 42*(2), 151–160.

Hawkins, R. D., Lalevic, N., Clark, G. A., & Kandel, E. R. (1989). Classical conditioning of the Aplysia siphon-withdrawal reflex exhibits response specificity. *Proceedings of the National Academy of Sciences, 86*(19), 7620–7624.

Heyes, C. (2017). When does social learning become cultural learning? *Developmental Science, 20*(2), e12350.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological bulletin, 136*(3), 390.

Holloway, R. L. (1967). The evolution of the human brain: Some notes toward a synthesis between neural structure and the evolution of complex behavior. *General Systems, 12*, 3–19.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological review, 109*(4), 679.

Holroyd, C. B., Larsen, J. T., & Cohen, J. D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology, 41*(2), 245–253.

Houwer, J. D., Baeyens, F., & Field, A. P. (2005). Associative learning of likes and dislikes: Some current controversies and possible ways forward. *Cognition and Emotion, 19*(2), 161–174.

Humphrey, N. K. (1976). The social function of intellect. In *Growing points in ethology* (pp. 303–317). Cambridge University Press.

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of personality and social psychology, 96*(5), 933.

Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The "how" question. In *Advances in experimental social psychology* (Vol. 43, pp. 205–255). Academic Press.*.

Kachergis, G. (2012). Learning nouns with domain-general associative learning mechanisms. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 34, No. 34).

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel prize lecture, 8*(1), 351–401.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., & Trope, Y. (2002). Reflexion and reflection: a social cognitive neuroscience approach to attributional inference.

LoTemplio, S. B., Scott, E. E., McDonnell, A. S., Hopman, R. J., Castro, S. C., McNay, G. D., … Strayer, D. L. (2020). Nature as a potential modulator of the error-related negativity: A registered report. *International Journal of Psychophysiology, 156*, 49–59.

Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science, 14*(1), 47–53.

Mackintosh, N. J., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science, 14*(1), 5–6.

Martin, I., & Levey, A. B. (1978). Evaluative conditioning. *Advances in Behaviour research and Therapy, 1*(2), 57–101.

Melloni, M., Lopez, V., & Ibanez, A. (2014). Empathy and contextual social cognition. *Cognitive, Affective, & Behavioral Neuroscience, 14*(1), 407–425.

Meyer, A., Carlton, C., Chong, L. J., & Wissemann, K. (2019). The presence of a controlling parent is related to an increase in the error-related negativity in 5–7 year-old children. *Journal of Abnormal Child Psychology, 47*(6), 935–945.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience, 24*(1), 167–202.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*(2), 183–198.

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology, 48*(2), 229–240.

Moore, J. W., Dickinson, A., & Fletcher, P. C. (2011). Sense of agency, associative learning, and schizotypy. *Consciousness and Cognition, 20*(3), 792–800.

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & emotion, 27*(4), 743–752.

Moran, T., & Bar-Anan, Y. (2020). The effect of co-occurrence and relational information on speeded evaluation. *Cognition and Emotion, 34*(1), 144–155.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60*, 157–162.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34*(5), 435–461.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology, 47*(6), 708–723.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences, 14*(10), 435–440.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38*(5), 752–760.

Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and self-regulation* (pp. 1–18). Boston, MA: Springer.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37*(4), 557–569.

Prados, J., Alvarez, B., Howarth, J., Stewart, K., Gibson, C. L., Hutchinson, C. V., … Davidson, C. (2013). Cue competition effects in the planarian. *Animal cognition, 16* (2), 177–186.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35*(2), 395–405.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of personality and social psychology, 91*(6), 995.

Scheffers, M. K., & Coles, M. G. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance, 26*(1), 141.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review, 84*(2), 127.

Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of experimental psychology, 38*(2), 168.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. Computational intelligence and neuroscience, 2011.

Tucker, D. M., Hartry-Speiser, A., McDougal, L., Luu, P., & Degrandpre, D. (1999). Mood and spatial memory: Emotion and right hemisphere contribution to spatial cognition. *Biological psychology, 50*(2), 103–125.

Umilta, C. (1988). The control operations of consciousness.

Van Bavel, J. J., & Cunningham, W. A. (2010). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European review of social psychology, 21*(1), 237–284.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological review, 111* (4), 931.

Yon, D., & Press, C. (2014). Back to the future: Synaesthesia could be due to associative learning. *Frontiers in psychology, 5*, 702.

Zanon, R., De Houwer, J., & Gast, A. (2012). Context effects in evaluative conditioning of implicit evaluations. *Learning and Motivation, 43*(3), 155–165.

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology, 67*(11), 2105–2122.