Human or not human? Performance monitoring ERPs during human agent and machine supervision

Bertille Somon, Aurélie Campagne, Arnaud Delorme, Bruno Berberian

NeuroImage xxx (2018) xxx-xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com

Human or not human? Performance monitoring ERPs during human agent and machine supervision

Bertille Somon[a, b, *], Aurélie Campagne[b], Arnaud Delorme[c, d], Bruno Berberian[a]

[a] ONERA, The French Aerospace Lab, Information Processing and Systems Department, 13661, Salon Cedex Air, France
[b] Univ. Grenoble Alpes, CNRS, LPNC, 38000, Grenoble, France
[c] Centre de Recherche Cerveau et Cognition, CNRS, Université Paul Sabatier, Toulouse, France
[d] Swartz Center for Computational Neuroscience, Institute of Neural Computation, University of California, San Diego, CA, USA

ARTICLE INFO

Keywords:
Supervision
Automation
Flanker task
Electroencephalography (EEG)
N2—P3 complex
Cluster-based permutation test

ABSTRACT

Performance monitoring is a critical process which allows us to both learn from our own errors, and also interact with other human beings. However, our increasingly automated world requires us to interact more and more with automated systems, especially in risky environments. The present EEG study aimed at investigating and comparing the neuro-functional correlates associated with performance monitoring of an automated system and a human agent using a vertically-oriented arrowhead version of the flanker task. Given the influence of task difficulty on performance monitoring, two levels of difficulty were considered in order to assess their impact on supervision activity. A large N2—P3 complex in fronto-central regions was observed for both human agent error detection and system error detection during supervision. Using a cluster-based permutation analysis, a significantly decreased P3-like component was found for system compared to human agent error detection. This variation is in line with various psychosocial behavioral studies showing a difference between human-human and human-machine interactions, even though it was not clearly anticipated. Finally, the activity observed during error detection was significantly reduced in the difficult condition compared to the easy one, for both system and human agent supervision. Overall, this study is a first step towards the characterization of the neurophysiological correlates underlying system supervision, and a better understanding of their evolution in more complex environments. To go further, these results need to be replicated in other experiments with various paradigms to assess the robustness of the pattern and decrease during system supervision.

1. Introduction

Our everyday life interactions with others rely highly on our ability to anticipate their actions, but also detect whenever they commit an error or not. This error detection process is called "performance monitoring". It is defined as "[…] a set of cognitive and affective functions determining whether adaptive control is needed and, if so, which type and magnitude is required" (Ullsperger et al., 2014). Other's performance monitoring is crucial in learning processes but also for correct social interactions (Tomasello et al., 1993). Even though the neural correlates of monitoring of our own errors are relatively well know (for a review on performance monitoring, see Gehring et al., 2011; Holroyd and Coles, 2002; Riesel et al., 2013; Taylor et al., 2007), the brain correlates associated with performance monitoring of others are still poorly documented (Bates et al., 2005; Koban et al., 2010; Van Schie, Mars, Coles and Bekkering, 2004). However, various studies suggest that similar brain processes are involved in supervision of another human and one's own performance monitoring (for reviews, see Ninomiya et al., 2018; and Somon et al., 2017).

During supervision tasks, several electroencephalographic (EEG) studies have shown that the detection of another person's error induces a negative component followed by a positive deflection in frontal and central regions. These components have been called the observation error-related negativity (oERN) and observation error positivity (oPe)(Carp et al., 2009; de Bruijn and Von Rhein, 2012; Koban et al., 2010; Van Schie et al., 2004; Weller et al., 2018) with regards to the Error-Related Negativity (ERN) and error Positivity (Pe) components described for our own performance monitoring (Falkenstein et al., 1991; Gehring et al., 1990). Moreover, functional magnetic resonance imaging (fMRI) studies have shown that the same brain regions seem to be activated during others' error and our own error detection (Cracco et al., 2015; Desmet and Brass, 2015; Jääskeläinen et al., 2016).

More recently, the increasing amount of automation technology in our daily-life has also raised questions about system performance monitoring.

* Corresponding author. ONERA, Base Aérienne 701, 13661, Salon Cedex Air, France.
Email address: bertille.somon@onera.fr (B. Somon)

https://doi.org/10.1016/j.neuroimage.2018.11.013
Received 18 May 2018; Received in revised form 23 October 2018; Accepted 9 November 2018
Available online xxx
1053-8119/

Adding automation has been considered for a long time as a simple substitution of human activity for machine activity (substitution myth, see Woods and Tinapple, 1999). However, automation has profoundly changed human activity at work. Human operators are now relegated to the role of passive supervisors: they are solely asked to monitor the actions and detect failures of automated systems (Moray, 1986; Sheridan, 1992, 1997; Sheridan and Verplank, 1978). This change in activity has led to new cognitive dysfunctions such as the out-of-the-loop performance problem (Endsley and Kiris, 1995; Kaber and Endsley, 1997). This issue, triggered by over-trust and complacency towards highly reliable automated systems, is characterized by an inability to detect errors correctly and to take over whenever necessary (Berberian et al., 2017). It has been pointed out as the cause of several accidents and incidents in aeronautics, for example (Shappell et al., 2007). A better understanding of the brain process related to the detection of others' errors, and particularly during system supervision, has thus become essential.

A few researchers have started to tackle the issue of performance monitoring during system supervision and have looked at its neural correlates. They assessed several types of system errors (Chavarriaga and Millán, 2010; Chavarriaga et al., 2014; Ferrez, 2007; Padrao et al., 2016; Shappell et al., 2007) or malfunctions (Desmet et al., 2014; Gentsch et al., 2009). Quite consistently, they were able to observe the same ERPs as those observe during human agent supervision (i.e., the oERN, oPe) at the same locations (i.e., Fz, Cz or FCz electrodes). fMRI studies have also shown that the same type of activity is observable, in the same brain regions as for human agent supervision (i.e., pMFC), when looking at system malfunctions. Nevertheless, several groups have identified another ERP during system supervision: the N400 (Padrão et al., 2016; Pavone et al., 2016). This component is usually linked to semantic processing, and more precisely to the observation of semantic aberrations, when recorded at the Pz electrode (Kutas and Hillyard, 1980). However, Balconi and Vitalloni (2014) have argued that this potential can be measured during the observation of aberrations in movement sequences at more frontal and temporo-parietal locations. Similarly, Padrão and colleagues (2016) and Pavone et al. (2016) observed this ERP during avatar error observation respectively at the Pz and FCz electrodes, whereas Ferrez and Millán (2005) observed it during Human-Machine Interaction (HMI) error detection at the Cz electrode. But this component seems to be observed solely whenever a movement, either of the avatar or of the HMI system, is involved.

To our knowledge, even though the neural correlates of performance monitoring seem similar, no direct comparison has been performed between human and artificial agent supervision while both executed the same task. No study has measured brain activity in reaction to supervision of an automated system performing a human task either. In addition, system error monitoring was mainly assessed through a few studies about system malfunctions triggered by human agent actions (Gentsch et al., 2009; Ullsperger et al., 2014; Padrão et al., 2016; Desmet et al., 2014) or by using human avatars (Pavone et al., 2016). Thus, they do not allow to specifically identify the possible effect of psychosocial parameters on performance supervision while several studies have shown that various psychosocial parameters differentiate interactions with human and artificial, automated agents. For example various emotions, feelings or representations can have an impact on performance monitoring ERPs activity. Interpersonal similarity (Carp et al., 2009), empathy (Marco-Pallarés et al., 2010), intentionality (Desmet and Brass, 2015), rivalry (Koban et al., 2010), agency feeling (Cracco et al., 2015), motor representation (Van Schie et al., 2004) also modulate the performance monitoring activity. Most of these processes cannot be developed with an automated system as they can with another human agent (see for example Riek, Rabinowitch, Chakrabarti and Robinson, 2009; Wohlschläger et al., 2003). Finally, Ninomiya et al. (2018) argue that observed errors in everyday tasks should have some kind of relevance for the observer in order to trigger the pMFC activity, thus the performance monitoring system. In comparison with literature, our study compares the supervision of both a human and an artificial agent with the same task configuration thus modulating only the psychosocial factors and controlling for any other factor which could modify the performance monitoring activity during supervision.

The present EEG study aims at characterizing the brain correlates related to performance monitoring of either a human agent or a system, in the same experimental paradigm. To this purpose, participants took part in a vertically-oriented arrowhead version of the flanker task in which they had to supervise another human and an artificial agent while he/it was performing the task. This experimental procedure allows to supervise both the human and artificial agents that perform exactly the same task thus modulating only the interaction type that the participant develops with each agent. Moreover, several studies suggest ERPs associated with performance monitoring in execution tasks are modulated by task difficulty (Van der Borght et al., 2016). However, the extent to which task difficulty modulates ERPs associated with performance supervision remains still largely unknown. Two levels of difficulty (easy and difficult) were thus considered in our study in order to assess the influence of task difficulty on supervision activity and to have a better understanding of ERPs' function in performance supervision tasks. The level of task difficulty was manipulated using distractors (difficult condition) or not (easy condition) above and below the target arrow in the modified flanker task. Both executor (the agent) and supervisor visualize the same type of stimulus for both easy and difficult conditions. In comparison with literature, this experimental procedure will allow to clarify the impact of distractors on the supervision activity. Indeed, in most studies measuring other's monitoring with the flanker task, the executor performs the equivalent of our difficult task (i.e. target with flankers), whereas the supervisor is only given the target (our easy task) as stimulus.

Concerning performance monitoring activity, we expect to observe fronto-central response-locked ERPs similar to that observed in the literature during one's own performance monitoring - i.e. observational ERN and Pe - for both human agent and system supervision. With regards to the effect of accuracy, irrespective of the agent type, the amplitude of the oERN and oPe are assumed to be higher for error supervision compared to correct response observation in accordance with the literature (Van Schie et al., 2004; Carp et al., 2009, de Bruijn and Von Rhein, 2012).

Concerning the influence of the agent type on performance monitoring, it remains under debate. Some authors suggest an increase of the performance monitoring activity for system supervision due to a decrease in interpersonal similarity, which is negatively correlated with the oERN amplitude (Carp et al., 2009). Conversely, a decrease of the oERN can be expected due to a decrease of the motor representation and intentionality for system compared to human agent supervision as both have been reported to impact supervision activity (Van Schie et al., 2004; Desmet and Brass, 2015). In our study, participants will not be faced with the other agent, nor presented thoroughly to him. So, based on the previous literature we can assume a decrease in performance monitoring activity as we believe that interpersonal similarity and motor representation won't have any effect in our study. Intentionality should have the highest effect thus decrease monitoring activity during system supervision compared to human agent supervision.

Concerning the effect of task difficulty on supervision, we expect a decrease of supervision ERPs amplitude with increasing level of task difficulty. The degree of uncertainty in the difficult condition of our study is higher due to the visual complexity of the stimulus and several studies have shown that amplitudes of performance monitoring components (ERN, Pe) were modulated by uncertainty and task difficulty (Van der Borght et al., 2016; Pailing and Segalowitz, 2004; Endrass et al., 2012). Consequently, we assume that the amplitudes of the oERN and oPe associated with errors should be decreased in the difficult condition compared to the easy condition. A similar effect of task difficulty on supervision ERPs is expected for human agent and system supervision, the task being similar for both agents.

Finally, in this study, in order to identify the brain activity specifically related to performance monitoring of another agent without preconceived notions about their spatiotemporal characteristics, the electrophysiological data were statistically analyzed using a robust cluster-based permutation test (Maris and Oostenveld, 2007). This analysis, which was at first introduced to investigate magnetic resonance imaging (MRI) data (Bullmore et al., 1999) does not require choosing a particular time window, neither a particular electrode location, to

formed on performance monitoring data. Results of this analysis will be compared to a more classical analysis of variance (ANOVA) at given electrode locations, in accordance with the literature and based on grand averages.

2. Materials and methods

2.1. Participants

A power analysis of performance monitoring data during supervision suggested a sample size of 7 subjects to detect both the oERN (de Bruijn and Von Rhein, 2012: 1-β = 0.80, $d_z = \frac{\sqrt{F}}{\sqrt{n}} = \frac{\sqrt{43.71}}{\sqrt{24}} = 1.350$, α = 0.05) and the oPe (Weller et al., 2018: 1-β = 0.80, $d_z = \frac{\sqrt{F}}{\sqrt{n}} = \frac{\sqrt{40.25}}{\sqrt{22}} = 1.353$, α = 0.05). Based on these results and the population sizes typically described in this literature in the research domain (including between 15 and 20 participants), we recruited seventeen healthy right-handed participants (12 men; 27.5 years ± 4.78 years) to perform the experiment from the general population via mailing-lists and published ads. Their laterality was measured with the Edinburgh inventory test (m ± SEM = 87.06 ± 4.54%; Oldfield, 1971). They had normal or corrected-to-normal vision and hearing, had no neurological or psychiatric disorders and were not under any medication. The study was approved by the local French ethics committee for non-interventional research (CERNI - Comité d'Ethique pour les Recherches Non Interventionnelles, IRB00010290-2016-09-13-12) of the Pôle Cognition Grenoble and conducted according to the principles expressed in the 1964 Declaration of Helsinki. A written informed consent was obtained from all participants, who received a financial compensation.

2.2. Experimental task and procedure

2.2.1. Stimuli

Task stimuli were displayed in white against a black background using the E-prime 2.0 software (v.2.0.10.356, E-prime Psychology Software Tools Inc., Pittsburg, USA) onto a 19-in CRT monitor (with a 1024 × 768 pixels resolution and a 100-Hz refresh rate) located 46 cm away from the participant in an unlit room. They consisted of five vertically-oriented arrowheads (2.8° × 0.6° of visual angle) that included a target (central arrowhead) and four flankers (2 arrowheads above and below the target). Two difficulty levels were considered (see Fig. 1). The easy condition only displayed the target arrowhead (0.5° × 0.6° of visual angle). It could either be pointing up, or down. The difficult condition displayed the target flanked above and below. The flanking arrowheads all pointed in the same direction, but could either be congruent with the target arrowhead (in the same direction) or incongruent (in the opposite direction).

Following the display of task stimulus, the agent's response according to the target orientation was displayed and consisted of the same arrowhead as presented in the easy condition.

2.2.2. Procedure

Participants took part in a modified version of the flanker task (Eriksen and Eriksen, 1974). They had to supervise and assess the accuracy of an artificial or a human agent in a modified vertically-oriented arrowhead version of the flanker task, using a response box (Chronos® Psychology Software Tools Inc., Pittsburg, USA). The experiment was divided into two difficulty sessions (easy and difficult), separated by at least a week. Sessions order was counterbalanced across participants: eight participants started with the easy condition, whereas the other nine started with the difficult one. Each session, lasting approximately 1 h, included ten task blocks performed by a human agent (a fellow coworker), and ten task blocks performed by an artificial agent (a computer), separated by breaks. The order of the 20 blocks was pseudo-randomized for each subject and differed between participants. However, the same order of blocks was used for both difficulty sessions. Participants were informed of the type of agent performing the task at the beginning of each block. In the easy session, each of the 20 blocks was composed of 72 trials (lasting 3.75 min): 36 with the target facing up and 36 with the target facing down pseudo-randomly presented. In the difficult session, each of the 20 blocks was composed of 48 trials (lasting 2.5 min) and 4 types of stimuli (congruent up and down, incongruent up and down) were equiprobable and pseudo-randomly presented.[1] Each trial started with the display of a fixation rectangle (4.85° × 1.9° of visual angle) for a variable duration (1 ± 0.25s), followed by the display of the task stimulus for 10 ms. A fixation point was displayed until the agent's – human or artificial – response (an arrowhead pointing up or down), which was then presented for 350 ms and followed by a jitter black screen for a duration of 300–350 ms. Each trial ended with the question "ERREUR?" ("ERROR?" in French) for 1 s and the participant had to state whether the agent was right or wrong by pressing the corresponding response key. Responses ("Oui" or "Non", i.e., "Yes" or "No" in French) associated with response buttons were counterbalanced across participants. Participants were previously familiarized with the task for each difficulty condition using trials performed by the artificial agent. No training trial was analyzed. Fig. 2 shows a complete description of a trial and stimuli.

In order to avoid any bias in the number of errors and correct responses performed by the other agent across participants, all trials were computerized. Still the fellow human agent stayed in the room next to the one of the participant during the whole experiment and came to see him at every break. The error rate for both sessions was set at 33.3%. Reaction times of the agent for each experimental condition were based on the reaction times of the participant obtained during a previous experimental session in which the same participants had to perform themselves the same modified version of the flanker task (data in revision).

2.3. Measure and analysis

2.3.1. Subjective and behavioral data

2.3.1.1. Subjective data Task difficulty was assessed at the end of each session, on a Likert scale from 0 to 10. Task difficulty was analyzed with a pairwise *t*-test with session type (easy vs. difficult) as within subject factor. Two participants did not fill the difficulty questionnaire at the end of the difficult session, and consequently were removed from the statistical analysis.

2.3.1.2. Behavioral data Participants' responses were recorded using the E-Prime 2.0 software (v.2.0.10.356) and analyzed with the R software (v.3.3.2, R Core Team, 2016). For each participant, error detection rates (EDRs) for agent accuracy were computed for each experimental condition (difficulty level, type of agent) as the ratio between the number of errors correctly detected by the participant and the total number of errors in the given condition (the total number of errors is 240 in the easy condition and 160 in the difficult condition, for each type of agent, per participant). EDRs were analyzed using a two-way repeated-measures ANOVA with agent type (human vs. artificial) and difficulty (easy vs. difficult)[2] as within subject factors. The *d'* coefficient was measured for each experimental condition. This measure corresponds to "…the detectability of a given signal for a given observer" (Swets et al., 1961). *d'* equals 0 if a participant obtains 50% accuracy (chance level) whereas a positive *d'* indicates a better than chance performance (Haatveit et al., 2010). *d'* values

[1] The difference in the number of trials between the easy and the difficult conditions comes from a pre-test phase that we performed with other participants in order to define several parameters of the task. We observed during this pre-test that the probability of making errors in the easy condition was lower than in the difficult condition. Thus we estimated that 72 trials were required in the easy condition to end up with the same number of errors than in the difficult condition, when performing the modified flanker task. This number of trials per difficulty condition was then validated during a prior session where participants had to perform themselves the modified flanker task, before the supervision task described here (data under revision). This number of trials was then kept in the supervisory task for both agent types.

[2] We performed a first analysis that took into account the congruency of the stimuli in the difficult conditions. We assessed the difference between EDRs with three difficulty levels: easy, difficult congruent and difficult incongruent. As there was no main effect of difficulty ($F(2,32) = 2.72$, $p = .08$), and for statistical robustness, we decided to collapse difficult congruent and incongruent trials for both behavioral and EEG data analysis.

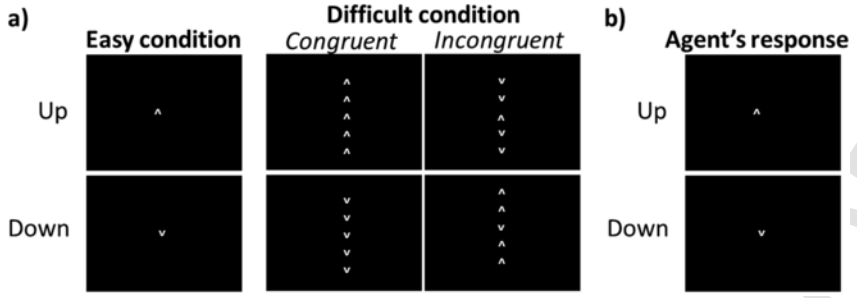

Fig. 1. Task stimuli (a) and agent's response stimuli (b) presented to the participant in the modified flanker task.

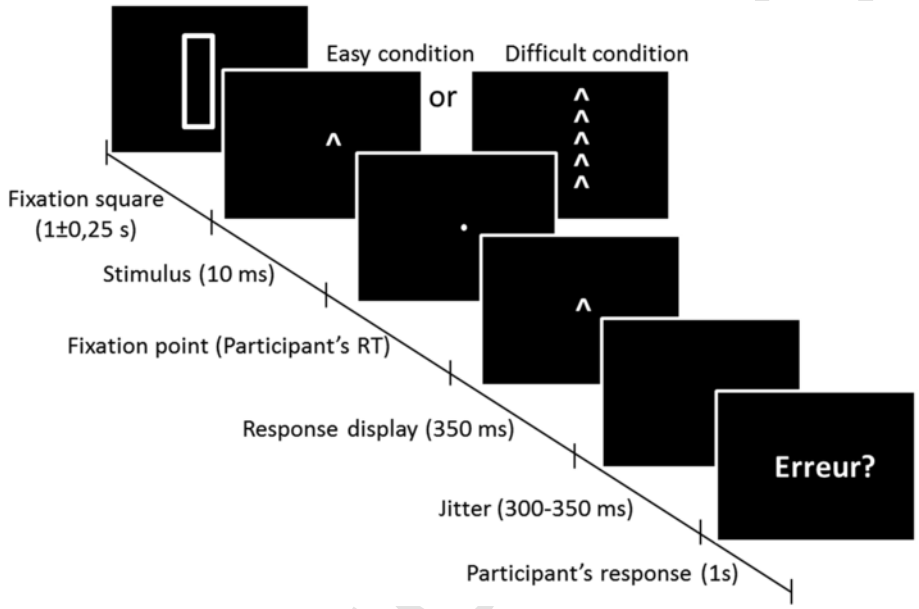

Fig. 2. Experimental design of the trial during the easy and difficult conditions of the supervised flanker task: participants had to determine whether the agent executing the task (human or artificial) performed correctly or not in the flanker task by comparing the orientation of the central target arrowhead of the task stimulus and the response given by the other agent.

were analyzed using a two-way repeated-measures ANOVA with type of agent (human vs. artificial) and difficulty (easy vs. difficult) as within subject factors.

For both EDRs and *d'* comparisons, partial eta squared was provided as a measure of the effect size and mean comparisons were performed using a Bonferroni correction. Reaction times were not analyzed, given participants were not required to respond as quickly as possible. All results are reported as *mean* ± *SEM*. The significance level was placed at .05.

2.3.2. Electroencephalography

The electroencephalogram (EEG) was continuously recorded using an actiCAP (Brain Products GmbH) equipped with 75 Ag/AgCl unipolar active electrodes (i.e., the 65 actiCAP montage to which we added F9, F10, P9, P10, PO9, PO10, O9, O10, M1, M2) which were positioned according to the extended 10–20 system (Oostenveld and Praamstra, 2001). The reference and ground electrodes used for EEG data acquisition were those of actiCAP and were positioned on the forehead (at AFz and Fpz electrodes respectively). Blinks and eye movements were also monitored using four pure silver electro-oculography electrodes: two positioned above and below the left eye on the median axis for vertical activities and two at the eyes' outer canthi for horizontal activities. The ground electrode for the EOG electrodes was placed on the earlobe. In addition, participants were instructed to limit blinking and eye-movements from the fixation point to their response. The signal impedance was kept below 10 kΩ for all electrodes. The signal was amplified using an actiCHamp system (Brain Vision, LLC), digitized at a 24-bit rate and sampled at 1,000 Hz, with a 0.05 μV resolution. No filtering was applied during data acquisition.

All EEG data analyses were performed using EEGLab (v.14.1.1; Delorme and Makeig, 2004) and Fieldtrip (Oostenveld et al., 2011) toolboxes on Matlab R2014b (v.8.4; The MathWorks, Inc.). Raw EEG data were re-referenced offline to the linked mastoids. The signal was segmented into 3-s epochs from trial onset. The signal was then down-sampled at 500 Hz and band-pass filtered between 0.5 and 30 Hz. All segments contaminated with muscular activity and/or non-physiological artifacts were rejected offline after a visual inspection. Artifacts related to ocular movements (saccades and blinks) were corrected using an Independent Component Analysis (ICA). Seventy-two independent components were utilized to perform the analysis, and on average 1.29 ± 0.11 components were removed per subject. Preprocessed epochs were segmented again from 200 ms before to 750 ms after the agent's response display in order to identify the event-related potentials (ERPs) time-locked to this response. Baseline correction was then applied from the 200 to 0 ms period preceding the agent's response display. For each participant, the ERPs induced by each type of agent's responses and measured over each electrode were averaged according to the accuracy of the response (agent error and correct response), the type of agent (human and system) and the level of task difficulty (easy and difficult). Trials which were misclassified by the participant (i.e. correct responses classified as errors, and vice versa) were excluded from EEG analyses.

Several ERPs were visually identified in fronto-central regions on basis of the grand averages of data in accordance with the literature. First, a positive wave (called P2) peaking at the FCz location was observed between 200 and 300 ms from the agent's response display (Pfefferbaum et al., 1985). This component was followed by a N2—P3 complex including a negative wave (N2) extending to 350 ms maximum, and then a positive wave (P3), which peaked at

the FCz location between 300 and 500 ms from the agent's response onset (Enriquez-Geppert et al., 2010; Gajewski and Falkenstein, 2013). The negative fronto-central activity, followed by a positive fronto-central one have been identified in the literature either as the observation ERN (oERN) and observation Pe (oPe) complex, or as the N2—P3 complex. The fact these two complexes represent the same activity is still under debate (Ullsperger et al., 2014). Here we chose to use the N2—P3 nomenclature in adequacy with the obtained data.

Peak amplitudes of the P2, N2 and P3 components were defined using the ERPLAB (v7.0) toolbox on Matlab R2014b (v8.4) and an adaptation of the local peak characterization function (with the "Neighborhood" parameter equal to 2; Lopez-Calderon and Luck, 2014).

Mean peak amplitudes of the P2, N2 and P3 components were classically analyzed at FCz with a three way repeated measures ANOVA with difficulty (easy vs. difficult), accuracy (error vs. correct) and agent type (human vs. artificial) as within subject factors. For robustness purposes of the EEG analysis, difficult trials were not separated considering the congruency of the stimulus. This choice was made for two main reasons: i) a first analysis on the EDRs considering congruency showed no main effect of difficulty on this measure ($F(2,32) = 2.72$, $p = .08$); and ii) data from a previous experimental session where the participants had to perform the same modified flanker task as with this study showed no effect of congruency on performance monitoring ERPs in the difficult condition ($F(1,16) = 0.97$, $p = .34$; data in revision). Mean comparisons were explored using Bonferroni post-hoc test (for multiple comparisons) and a significance threshold set at 0.05.

In order to identify brain activities specifically related to performance monitoring of another agent, without preconceived notions about their spatiotemporal characteristics, the ERP data were also statistically analyzed using the robust cluster-based permutation test (Maris and Oostenveld, 2007) with the Fieldtrip toolbox on Matlab R2014b (v.8.4). This analysis is based on the cluster mass test (Oostenveld et al., 2011) which was at first introduced to investigate MRI data (Bullmore et al., 1999).

This method identifies spatio-temporal clusters, in ERP data, presenting a significant difference between the conditions (accuracy, difficulty and type of agent as factors) in a given time window. With the cluster-based permutation test, two-dimensional EEG data (spatial and temporal dimensions) are averaged for each condition and for all the participants. There are two main steps for this analysis. The first step consists in selecting the significant clusters. To begin with, for every channel x time-point pair (a sample), the experimental conditions are compared two by two with a *t*-test. Each sample for which the test statistic is larger than a predefined threshold (here the value of the *t*-test statistics for $\alpha = .05$) is selected in a subset of samples. Then, clusters are drawn from this subset according to the temporal and spatial adjacency of the samples, but also to similarity in sign and magnitude. Here, we consider a minimum of 2 neighboring electrodes per cluster and a minimum duration of 20 ms. Finally, cluster statistics are calculated by summing the t-values of each sample in the cluster (definition of observed statistics: t_{obs}). This leads to the second step of the analysis: the permutation test. For this non-parametric statistical analysis, the averages for the two experimental conditions compared are first randomly assigned to 2 subsets of data for every subject. These subsets are called a random partition. Second, the samples in these random partitions are compared two-by-two. Third, clusters are drawn from these random partition sample statistics, with the same spatio-temporal constraints than in the observed cluster selection. Fourth the largest cluster is selected as the one for which the sum of samples statistic is the maximum. Fifth, another random partition is created and analyzed, and so on. A total of five hundred random partitions of data were computed and analyzed using the same scheme. This repetition yields a non-parametric Monte-Carlo randomization procedure to estimate the empirical distribution of the test statistic under the null hypothesis: each t-value is used to construct a histogram. Sixth, p-values are finally calculated for each cluster from the test statistic that was actually observed and the histogram as the proportion of random partitions that resulted in a larger test statistic than the observed one. A more visual description of these steps is provided in the Supplementary materials (see Fig. S1).

ERP data in the various conditions are then considered for each cluster by averaging across channels included in the cluster. The differences between conditions are now evaluated through a single test statistic for the complete grid of spatio-temporal pairs. The cluster-based permutation analysis revealed several significant clusters which are described in the results section.

3. Results

3.1. Subjective and behavioral data

3.1.1. Difficulty assessment

Participants reported that the task was harder to perform in the difficult condition (4.60 ± 0.58) compared to the easy one (3.35 ± 0.46 on a scale from 0 to 10, $t(14) = 3.073$, $p < .01$).

3.1.2. Error detection rates (EDRs)

EDRs were not modulated by the type of agent ($F(1,16) = 0.78$, $p = .39$) nor by task difficulty ($F(1,16) = 0.48$, $p = .5$; see Fig. 3a).

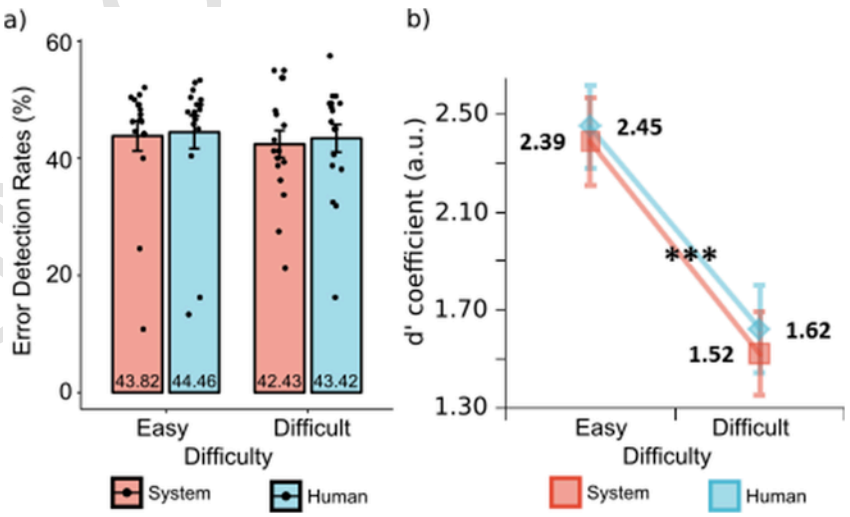

Fig. 3. Mean Error Detection Rate (a) and Detectability of stimuli (b) for all easy (left) and difficult (right) conditions, for human (red) and artificial (blue) agent supervision. Black points show the individual values for each participant in every condition. Error bars show standard errors to the mean (SEM) a.u.: arbitrary unit. ***: $p < .005$. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.1.3. d' value

Overall *d'* value was 2.20, and was significantly above chance performance ($F(1,16) = 204.02$, $p < .001$, $\eta^2_P = .93$). No effect of the type of agent ($F(1,16) = 2.11$, $p = .17$) and no interaction effect with this factor ($F(1,16) = 0.12$, $p = .73$) were observed on the values of *d'*. *d'* was only significantly modulated by task difficulty ($F(1,16) = 17.24$, $p < .001$, $\eta^2_P = .52$). Mean comparisons revealed that *d'* was significantly higher in the easy condition than in the difficult condition ($p < .005$, see Fig. 3b).

3.2. Monopolar ERPs

3.2.1. P2 component

The classical statistical analysis revealed no effect of agent type ($F(1,16) = 0.38$, $p = .55$) and no interaction of other factors with this factor on the P2 amplitude. The P2 amplitude was however significantly modulated by accuracy ($F(1,16) = 4.97$, $p < .05$, $\eta^2_P = .24$), task difficulty ($F(1,16) = 5.24$, $p < .05$, $\eta^2_P = .25$) and tended to be significantly modulated by the difficulty × accuracy interaction ($F(1,16) = 4.21$, $p = .057$, $\eta^2_P = .21$; see Fig. 4). Mean comparisons revealed that the P2 ERP was higher for detected errors compared to detected correct responses and in the easy compared to the difficult condition (see Fig. 4a and b).

3.2.2. N2 component

The classical statistical analysis revealed no effect of agent type ($F(1,16) = 3.7 \times 10^{-3}$, $p = .95$) and no interaction with this factor on the N2 amplitude. The amplitude of the N2 component tended to be modulated by accuracy ($F(1,16) = 4.40$, $p = .052$, $\eta^2_P = .22$) and was significantly modulated by task difficulty ($F(1,16) = 4.60$, $p < .05$, $\eta^2_P = .22$) and the difficulty × accuracy interaction ($F(1,16) = 4.61$, $p < .05$, $\eta^2_P = .22$, see Fig. 4). Mean comparisons showed that the N2 amplitude was significantly higher (lowest value) in the difficult compared to the easy condition. The interaction was revealed in the fact that the N2 amplitude following error detection in the difficult condition was significantly higher than for all the other conditions (all $p < .005$, see Fig. 4a and b). No difference in N2 amplitude was observed between errors and correct responses in the easy condition, neither between the easy and difficult conditions for correct responses.

3.2.3. P3 component

The classical statistical analysis revealed no effect of agent type ($F(1,16) = 4.6 \times 10^{-3}$, $p = .95$) and no interaction with this factor on the P3 amplitude. The amplitude of P3 was significantly modulated by accuracy ($F(1,16) = 69.42$, $p < .005$, $\eta^2_P = .81$), task difficulty ($F(1,16) = 15.77$, $p < .005$, $\eta^2_P = .50$), and the difficulty × accuracy interaction ($F(1,16) = 11.07$, $p < .005$, $\eta^2_P = .41$, see Fig. 4). Mean comparisons revealed that the P3 amplitude was significantly higher for detected errors than for detected correct responses in both easy and difficult conditions (both $p < .005$, see Fig. 4a and b). The interaction effect was reflected in the fact that the easy condition induced a significantly higher P3 amplitude than the difficult condition for detected errors only ($p < .005$).

3.3. Cluster-based permutation test

For main effects, we observed two clusters which were significantly different between errors and correct responses detection, regardless the type of agent and the difficulty level: i) one associated with a negative potential in the right central region (36 electrodes activated) from 160 to 340 ms post-agent response ($p < .05$), and ii) one associated with a positive potential in a large fronto-centro-parietal region (all electrodes activated except TP9, P10, PO10, O9/O10) from 340 to 750 ms ($p < .005$). Both clusters were higher for errors than for correct responses. The corresponding ERPs are assumed to be the N2 – for the

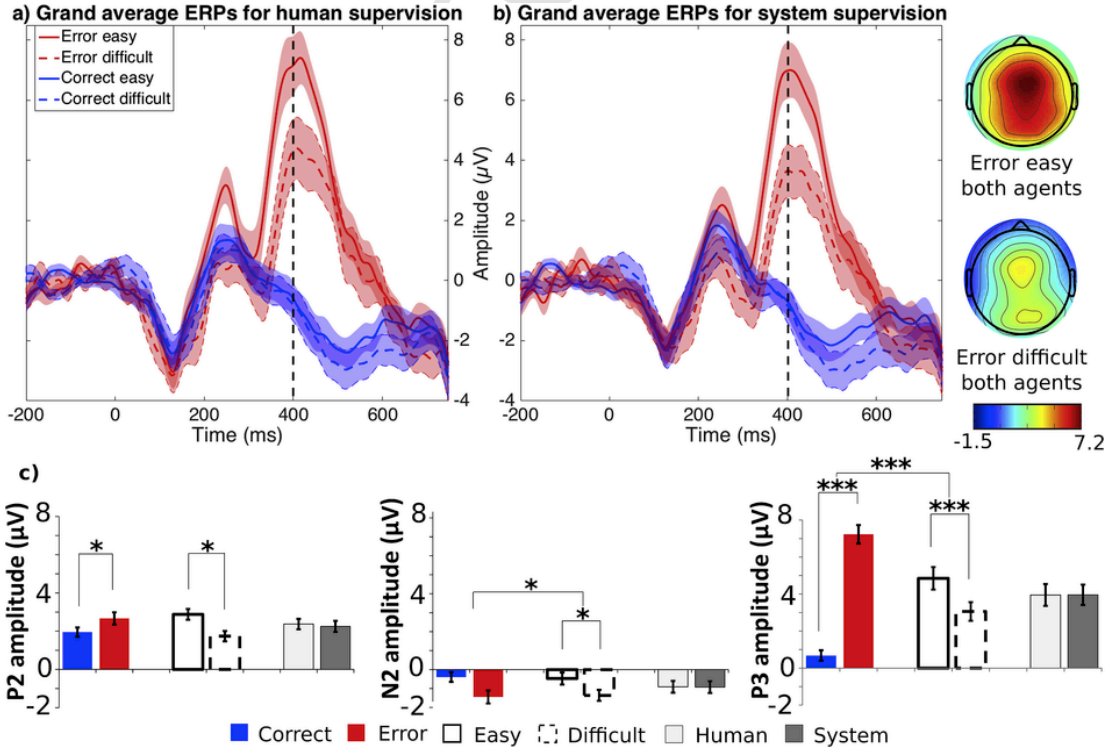

Fig. 4. Time-course of event-related potentials time-locked to the agent's response display (0 ms) at the FCz electrode for erroneous responses (red) in the easy (plain line) and difficult (dashed line) conditions, and for correct responses (blue) in the easy (plain line) and difficult (dashed line) conditions, for (a) human agent supervision and (b) system supervision. Waveforms are represented as mean ± SD across participants for each condition. The far right panel represents the topographies at the time-point 410 ms post-response (black vertical dashed line on the graphs) for errors in the difficult and easy condition with both human and system supervision responses averaged. (c) Results of the statistical analysis are represented for the P2 (left), N2 (middle) and P3 (right) components. Amplitudes give mean ± SEM for every category of the ANOVA (Accuracy, Difficulty and Type of Agent). *: $p < .05$; ***: $p < .005$. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

negative potential – and the P3 – for the positive potential. Concerning the effect of agent, no significant cluster was obtained. Concerning the effect of difficulty, a significant cluster associated with positive activity was observed in the centro-parietal region (44 electrodes activated) from 0 to 110 ms which was significantly higher in the difficult, compared to the easy condition. This cluster is assumed to correspond to the P2 component.

We then looked deeper into the comparisons of the conditions two-by-two. Only the significant comparisons are presented hereafter.

For human agent supervision, the comparison between correct responses and error detection based on permutation analysis revealed one significant cluster in a large fronto-centro-parietal region associated with a positive potential: i) in the easy condition (all electrodes activated except for P10, PO10, O9/O10) from 330 to 750 ms post agent response ($p < .005$), and ii) in the difficult condition from 360 to 750 ms post agent response with the same topography than in easy condition ($p < .005$). The positive potential was significantly higher for detected errors than for detected correct responses in both the easy and difficult conditions (see Fig. 5a).

For artificial agent supervision, the comparison between correct responses and error detection based on permutation analysis revealed one significant cluster in a large fronto-centro-parietal region associated with a positive potential, similar to the cluster observed for the human agent: i) in the easy condition (all electrodes activated except for TP9, P10, PO10, O9/O10) from 340 to 750 ms post agent response ($p < .005$), and ii) in the difficult condition (all electrodes activated except for TP10, P10, PO10, O9/O10) from 350 to 750 ms post agent response ($p < .005$). For both difficulties, the positive potential elicited was significantly higher for detected errors than for detected correct responses (see Fig. 5b). This result was similar to the ones obtained for human agent supervision. They are assumed to correspond to the P3 component of the N2—P3 complex observed on grand average ERPs.

For artificial agent supervision and in the difficult condition only, an accuracy effect was also observed on another cluster, located in right central regions and associated with a negative potential from 240 to 370 ms post agent response ($p < .05$). The permutation test revealed that this negative potential was significantly higher for detected errors than for detected correct responses (see Fig. 6a). This cluster can be assimilated to the N2 component of the N2—P3 complex.

The comparison between human and artificial agent supervision based on permutation analysis according to accuracy of responses revealed a significant cluster in fronto-centro-parietal regions associated with a positive potential from 400 to 500 ms post agent response ($p < .05$). This significant cluster was only observed for error detection, regardless of task difficulty (both difficulty levels grouped). It is assumed to correspond to the P3 component of the N2—P3 complex (see Fig. 6b for its time course and topography). This P3 component was significantly higher for human error detection than for system error detection.

The comparison between the easy and difficult conditions based on permutation analysis according to accuracy of responses revealed a significant cluster in fronto-central regions associated with a positive potential from 180 to 590 ms post agent response ($p < .005$). This significant cluster was only observed for error detection for both human and artificial agents grouped. This potential, assumed to correspond to the P2 and P3 components, was significantly higher in the easy condition than in the difficult condition.

4. Discussion

The main aim of this EEG study was to characterize ERPs related to performance monitoring during supervision of either a human or an artificial agent. To this purpose, participants took part in a modified vertically-oriented arrowhead version of the flanker task. In order to assess the influence of task difficulty on supervision activity, two levels of task difficulty were considered. Finally, brain activities specifically related to performance monitoring of another agent were statistically analyzed (i) with classical analyses of variance on peak amplitude of grand average ERPs and (ii) without preconceived notions about their spatiotemporal characteristics using the robust non-parametric cluster-based permutation test. Three main questions were asked in this study: i) Is it possible to observe performance monitoring activity in a supervision context, i.e. a cerebral activity triggered by error detection? ii) If so, is this error-related activity different when we supervise an automated system as compared to a human agent? iii) Does the difficulty of the task have an impact on the brain activity linked to other agent supervision? Thus, we will now discuss our results with regards to these three questions.

4.1. Supervision ERPs of others' performance

Although error detection during supervision has attracted more and more attention, it is not as well documented as error commission studies (Riesel et

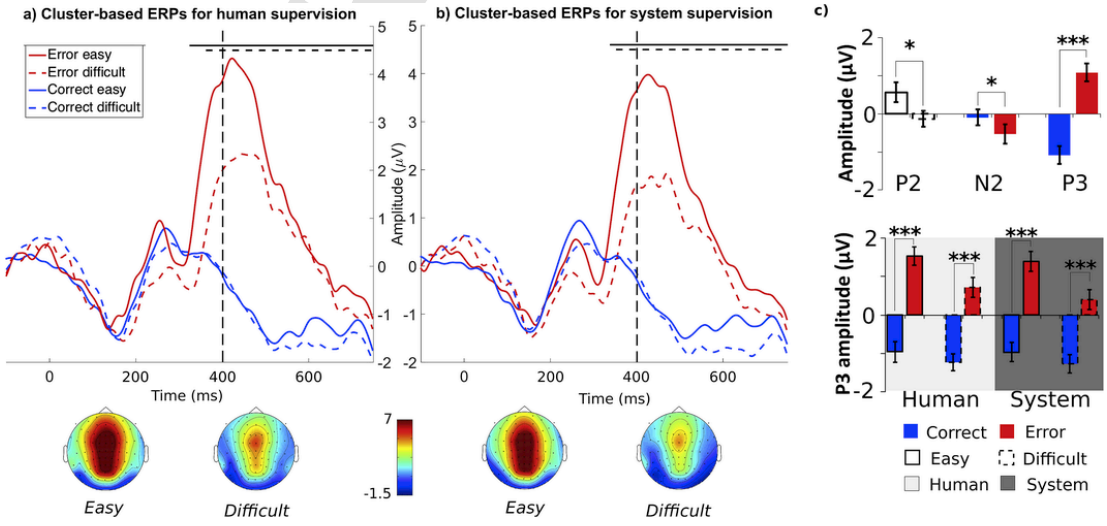

Fig. 5. Time course of significant clusters obtained with cluster-based permutation analysis on EEG data after error (red) and correct response (blue) detection in the easy (plain line) and difficult (dashed line) conditions (a) Significant differences for human agent supervision in the easy (horizontal plain line) and difficult (horizontal dashed line) conditions. The bottom panel represents topographies at time-point 410 ms (black vertical dashed line on the graphs) for the cluster-level difference wave between human agent detected errors and correct responses in the easy (left) and difficult (right) conditions for the cluster's electrodes only. (b) Significant differences for artificial agent supervision in the easy (horizontal plain line) and difficult (horizontal dashed line) conditions. The bottom panel represents topographies at time-point 410 ms (black vertical dashed line on the graphs) for the cluster-level difference wave between system detected errors and correct responses in the easy (left) and difficult (right) conditions for the cluster's electrodes only. (c) Results for the statistical analyses are presented for significant clusters only as the mean ± SEM across participants on the significant period of time (results not presented correspond to non-significant differences, thus resulting into no significant cluster). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

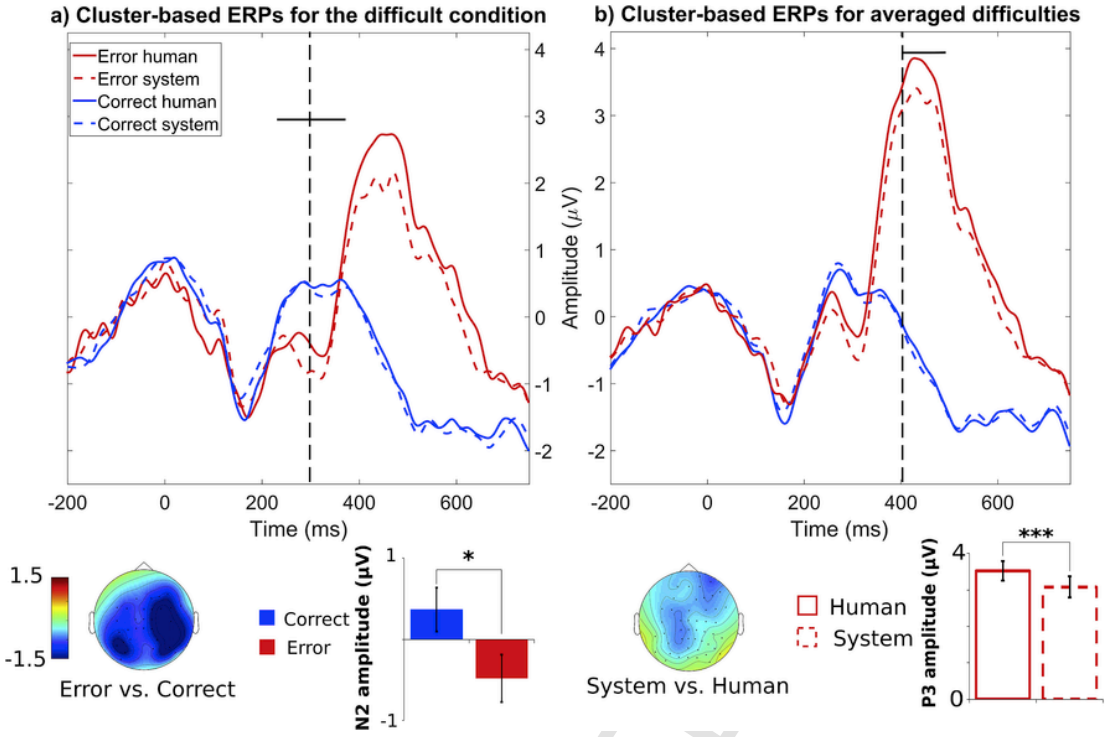

Fig. 6. Time course of significant clusters obtained with cluster-based permutation analysis on EEG data after error (red) and correct response (blue) detection for human agent (plain lines) and system (dashed lines) supervision (a) Negative cluster (i.e., N2 component). in the difficult condition significant for system error detection only (horizontal plain line). The topography (bottom panel) shows the differential activity between the two significantly different conditions at time-point 300 ms post-system response (black dashed vertical line) in the difficult condition for the cluster's electrodes only. Bar graphs show the statistical results for this significant cluster as mean ± SEM across participants on the significant period of time (b) Significant difference (horizontal plain line) between human agent and system error detection on the P3 component for both difficulties averaged. The topography (bottom panel) shows the differential activity at the time-point 410 ms post-errors (black dashed vertical line) for the cluster's electrodes only. Bar graphs show the statistical results for this significant cluster as mean ± SEM across participants on the significant period of time. (Results not presented on the bar graphs correspond to non-significant differences, thus resulting into no significant cluster). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

al., 2013; Somon et al., 2017). Our study gives new insights into the characterization of the performance monitoring activity during supervision. Indeed, we observed that there was a specific brain activity related to error detection during both human and artificial agent supervision. This time-locked activity shares the same characteristics as a well-known ERP complex: the N2—P3 complex. It was observed in its usual time window (i.e. 200–750 ms post-stimulus) with a maximum at the FCz electrode. This complex was coupled with a P2 component, significantly higher for errors than for the correct responses of the supervised agent. Several supervision studies (see e.g., Carp et al., 2009; de Bruijn et al., 2007; Ferrez and Millán, 2005, 2008), using flanker tasks, action slips or human-machine interfaces, revealed the same broad complex as that obtained in our study following error observation with similar topographies and latencies: a fronto-central positivity, followed by a central negativity and a fronto-central positivity. Ferrez and Millán (2005, 2008) actually called this complex the "interaction Error-related Potential" (iErrP) during performance monitoring of a HMI. Other studies have argued that they detected the "observational" counterpart of the ERN and Pe (oERN and oPe; Bates et al., 2005; Koban et al., 2010; Van Schie et al., 2004) and thus identified respectively a negative component followed by a positive one after error detection at fronto-central sites. The components observed in these studies tend to differ from those obtained in our studies in their shape and time course but these differences can be justified by the use of other tasks (Go-NoGo task) and different supervision contexts (e.g. cooperation/competition, different positions of the observer, difference in the difficulty of stimuli presented to the performer and the observer). Interestingly, several studies have suggested that the N2—P3 complex, observed in Go-NoGo tasks for example, could correspond to the same cognitive process as the ERN-Pe complex observed in performance monitoring/error commission tasks. In this direction, Ullsperger et al. (2014) recently made the hypothesis that the ERN, N200 and FRN were three representations of a unique process occurring at different stages of goal-directed behavior. Based on both classical ERP results and computational models, they argued that these three ERPs would reflect a fast alarm signal responding to an eliciting event (stimulus or action) occurring before, during or after the action. Likewise, they argue that the late Pe and P3b following these ERPs both represent the same subjective evidence signal. Our results tend to back up Ullsperger and colleagues' theory and thus suggest that the N2—P3 and ERN-Pe do both correspond to the same kind of performance monitoring activity.

The cluster-based permutation analysis revealed a similar pattern of results, with an effect of accuracy. Nevertheless, the cluster-based test gives more informative results. On our data, the algorithm identified significantly larger deflections for error detection compared to correct response detection from 160 to 750 ms after the supervised agent's response: first negative (between 160 and 340 ms) then positive (between 340 and 750 ms). The negative cluster included 35 electrodes in the right central region. This cluster was considered to correspond to the N2 component. The identified positive cluster was very wide as it included more than 65 electrodes in fronto-centro-parietal regions, regardless task difficulty and supervised agent type. This deflection was considered to be the P3 component. Although the P3 component is usually measured only at a few arbitrarily selected number of electrodes (at the location of the peak of the component) and as the mean in a time-window ranging from 300 to 500 ms, the cluster-based permutation analysis shows that this component is statistically broader as all the spatial and temporal points statistically significantly different between the two conditions are considered. The statistical scale is much higher and we can see it on the statistically relevant 67 electrodes and more than 200 time-points for the P300. This result was also observed when subdividing the data according to the type of agent: the P3 cluster was observed for error detection for human agent supervision, but also for system supervision, separately. These complementary findings result from the cluster-based permutation test's properties. Indeed, this technique allows to observe significant differences between various experimental conditions without defining a priori the localization or approximate starting time point of such activity (Maris and Oostenveld, 2007). It is also a non-parametric analysis (i.e., it requires no assumptions on

the data). Another benefit of this technique is that it takes into account the location and proximity of the electrodes, defining clusters of close electrodes.

In addition, in the difficult condition and for system supervision only, a significant cluster associated with a negative fluctuation was also measured in right central regions between 240 and 370 ms post-response when comparing the detection of errors and correct responses. The amplitude of this activity was significantly higher for error detection than for correct responses detection. Given its polarity, latency, and topography, this cluster was assumed to correspond to the N2 component peaking at the FCz electrode in grand averages. The frequency theory (Donkers and van Boxtel, 2005) can nicely explain such a difference appearing only in the difficult condition, and for system supervision. Indeed, this theory assumes that the N2 amplitude is modulated by the frequency of stimuli. Particularly, the N2 component is increased when faced with an attended deviant stimulus: its amplitude appears to be negatively correlated with the frequency of this deviant stimulus. In other words, the more the deviant stimulus is rare (the least frequent), the greater the amplitude of the N2 associated with it (Enriquez-Geppert et al., 2010). In our experiment, the error rate was equivalent for both difficulty levels (33.3%) but the actual detectability of error stimuli, as measured by the d' coefficient, was significantly smaller in the difficult condition compared to the easy condition. Thus, errors in the difficult condition represent the least frequent – or the most deviant – stimuli, which could explain why the amplitude of the N2 is larger in this condition. Similarly, the performance of the two agents was equivalent, but the verbal reports collected at the end of the experiment indicate that participants perceived the errors as less frequent for system than for human supervision (see next section), thus explaining why the amplitude of the N2 is larger for system error detection in the difficult condition. Interestingly, this latter result illustrates an impact of both agent type and task difficulty on the brain components associated with the performance monitoring of another agent. These impacts are detailed and discussed below.

4.2. Error detection and type of agent

The ANOVA on peak amplitude in grand averages revealed no main effect of the type of agent supervised, nor any interaction effect with this factor, whatever the ERPs considered (N2, P2 and P3). On the other hand, the cluster-based permutation analysis revealed a cluster significantly modulated by the type of agent and associated with a positive deflection in fronto-centro-parietal regions from 400 to 500 ms post-erroneous response of the supervised agent, corresponding to the time course and location of the P3 wave. This positive deflection was significantly lower for automated system than for human agent error detection, regardless task difficulty (i.e. when both difficulties were grouped). This result again shows the interest of the cluster-based permutation test compared to classical ANOVA. Indeed we can see on the data that the amplitude difference is very small. Thus the analysis of variance might not have picked it up. But the permutation test, through the Monte Carlo estimation allowed to fit more the data. Interestingly, some participants (11) filled a questionnaire after the supervision task.[3] All reported that the supervision of the human agent was different from the supervision of the automated system. Even though the error rate was similar and error presentation was randomized for the two types of agents, seven participants out of the eleven stated spontaneously that the human agent tended to either make more errors, or to make series of errors more than the artificial agent. In the performance monitoring literature, human agent supervision has been studied more than system supervision (except for BCI errors detection). The comparison between both has never been performed to our knowledge. Nevertheless, alike in our results, separate studies have shown error detection activity for both human agent and system supervision.

Several assumptions may justify this difference between human agent and system supervision in our study. Various studies have shown that human-human interactions differ from human-system interactions. A first assumption refers to the impact of similarity with the observed agent. In a study assessing error observation, Carp and collaborators (2009) showed that interpersonal similarity, as measured based on participants' beliefs and opinions, had an impact on brain activity associated with supervision of another human agent. They showed that the interpersonal similarity of the participant was both negatively correlated with the oERN amplitude and positively correlated with the oPe amplitude. Further, Riek et al. (2009) showed in a study that increase in "humanity" of a robot improved the participants' empathy for this robot. In a neuroimagery study, Shane and collaborators (2009) observed that empathic concern during other's performance observation in a Go/NoGo task modulates the activity in the rostral/ventral ACC (part of the medial prefrontal cortex). This result was backed up by another study by Newman-Norlund and collaborators (2009), but also by studies on pathological populations (Fitzgerald et al., 2005). All similarity effects are likely candidates to explain the decrease of the amplitude of performance monitoring ERPs when supervising the artificial compared to the human agent in our study.

Another assumption involves trust. The role of trust in human-automation interaction has been the focus of much research over the past decade (e.g., Dzindolet et al., 2003; for a comprehensive review, see Lee and See, 2004; Madhavan and Wiegmann, 2007). Particularly, it has been proven that high levels of automation could lead to over-reliance and failure to monitor the "raw" information sources provided as input to automation – the so-called complacency effect (Moray and Inagaki, 2000; Sheridan and Parasuraman, 2005). Yet, Lewandowsky and colleagues (2000) observed that human operators find automated systems more trustworthy than human collaborators when performing a task allocation work. Moreover, increasing complacency is often associated to decreasing attention. Several studies have assessed the effect of attention, or attention allocation, on the P300 amplitude and showed that it was greater for attended stimuli compared to unattended ones (Donchin and Coles, 1988; Johnson and Donchin, 1978). An increase of trust and complacency toward the automated system may thus explain our results, i.e. a lower positive cluster amplitude, assimilated to the P3 component, for system compared to human agent supervision.

Finally, the concept of intentionality attribution could also justify the difference observed between human and artificial agent supervision. Indeed, at the behavioral level Wohlschläger et al. (2003) showed that it is more difficult to attribute intentionality to an artificial agent than it is to another human agent, as measured by intentional binding. In neuroimagery, Desmet and Brass (2015) showed that the intentionality and usuality of an action performed by someone else modifies the activity in the medial pre-frontal cortex (MPFC). They observed an antero-posterior gradient of activation in the MPFC: the posterior MPFC was more activated for unusual accidental actions (i.e., errors) than for unusual intentional actions, and vice versa for the anterior MPFC. They also observed increased activation of other areas implicated in performance monitoring (e.g., anterior insula, inferior frontal gyri) for accidental actions observation compared to intentional ones. Thus a decrease of intentionality decreases the activity in the anterior MPFC which could also explain the decrease in the positive cluster amplitude assimilated to the P3 component for system supervision compared to human agent supervision.

4.3. Supervision and task difficulty

The ANOVA revealed a lower amplitude of the P3 component in the difficult condition compared to the easy condition, for error de-

[3] They were proposed 3 scales: i) to state which type of agent was more difficult to supervise (scale from +5 – human – to −5 – computer), ii) to state their confidence level towards the automated system (from 0 – no confidence – to 10 – confident), iii) to state their confidence level towards the human agent (from 0 – no confidence – to 10 – confident). They were also asked, as an open question, whether they observed any difference between the human agent and the automated system. Unfortunately the too small number of participants who filled the questionnaires did not allow us to report statistical results on these data. But the 11 subjects interviewed (6 for both sessions, 5 for one session) all reported at least one difference between the two types of agents. On average, they reported that both the human agent and the automated system were as easy to supervise (m ± SEM: 0 ± 0.16). Mean confidence towards the automated system was 6.24 ± 0.36; and mean confidence towards the human agent was 6.06 ± 0.36. Finally, 4 participants out of 11 reported that a human agent error led more to a succession of errors, 2 participants reported that the human agent made more errors than the automated

task difficulty also appears stable and robust as similar results were observed using the cluster-based permutation analysis. Although the role of the P2 and P3 components remains under debate, a majority of studies suggests a close link with attentional processes. Our results support the theories according to which the P3 amplitude variations reflect the stimuli categorization processes or target salience (Kök, 2001). Indeed, the P3 amplitude decreases with increasing categorization uncertainty, and with decreasing salience of the target. Both those factors are modulated in our task. The introduction of flankers in the difficult condition decreases the salience of the target compared to the easy condition. Likewise, as the salience of the target is decreased, the categorization uncertainty is increased in the difficult condition, compared to the easy one. Thus the impact of difficulty on our data is consistent with the literature.

In addition, the cluster-based permutation analysis reveals a main effect of difficulty on the P2 cluster from 0 to 110 ms. This analysis also allowed to identify a large cluster in fronto-central regions (61 electrodes) associated with a significantly larger positive deflection in the difficult condition than in the easy condition as early as 180 ms and up to 590 ms after the response of the supervised agent for only error detection and for both human agent and system. This deflection can be considered to include the P2 and P3 components.

An important aspect of our study is that the stimulus is the same for the participant and the supervised agent for both difficulty levels. Indeed, in everyday-life situations, the supervisor must be able to determine whether the agent's response is correct or erroneous based on his own analysis of stimuli that may be more or less complex. Nevertheless, at a theoretical level, our manipulation doesn't allow to determine whether the difference in brain activity between the easy and difficult conditions is due to a difficulty of the task at hand per se, or if it is due to differences in supervision difficulty. This aspect is also relevant as (i) distinguishing the impact of both allows to understand more the supervision and performance monitoring process, and (ii) it might be of interest in everyday-life situations where the supervision activity may be degraded because of various parameters. This question was tackled by a few researchers at the execution level (Scheffers and Coles, 2000) and could also be studied for the supervision activity in future works, in continuity of our research.

5. Conclusion

The error detection process takes place in our everyday-life, when we are performing various actions, but also when we observe or supervise the actions of another human agent or a system. The results of our EEG study revealed that the detection of errors performed by another agent or a system was characterized at the cerebral level by a larger P2—N2—P3 complex than the detection of correct responses in an extended fronto-centro-parietal region. Using a cluster-based permutation analysis, a lower positive fluctuation, considered as the P300, in fronto-centro-parietal regions was found for system supervision compared to human agent supervision. Furthermore, task difficulty only impacted error detection and modulated the entire P2—N2—P3 complex, for both human agent and system supervision. Better characterizing neurophysiological correlates underlying supervision will help better understand the associated cognitive dysfunctions that may appear with increasing automation and in degraded conditions. A replication of these results with other experimental paradigms would allow to determine if this pattern holds. Monitoring difficulties, like the "out-of-the-control-loop phenomenon", have been characterized at the behavioral level, but have rarely been looked at from a neuroscientific perspective. Our study can help apprehend how real-time control, supervision and error detection processes can be degraded when interacting with highly automated systems in our everyday life, and what the effects are on the brain.

Ethical statement

The study was approved by the local French ethics committee for non-interventional research (CERNI - Comité d'Ethique pour les Recherches Non Interventionnelles) of the Pôle Cognition Grenoble[4] and conducted according to the principles expressed in the 1964 Declaration of Helsinki. A written informed consent was obtained from all participants.

Funding

This work was supported by the 'Région Provence-Alpes-Côte d'Azur' (Emploi Jeunes Doctorants – 2015_07459) and the ANR/FRAE (Young researcher program – ANR-15-CE26-0010-01).

Acknowledgement

We would like to acknowledge Nicolas Maille for his much appreciated help during the experimentations and Sylvain Harquel for his informed recommendation concerning data analysis.

We also thank "AWS traduction" for their proofreading services.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2018.11.013.

References

Balconi, M., Vitaloni, S., 2014. N400 effect when a semantic anomaly is detected in action representation. A source localization analysis. J. Clin. Neurophysiol. 31 (1), 58–64 https://doi.org/10.1097/WNP.0000000000000017.

Bates, A.T., Patel, T.P., Liddle, P.F., 2005. External behavior monitoring mirrors internal behavior monitoring: error-related negativity for observed errors. J. Psychophysiol. 19 (4), 281–288 https://doi.org/10.1027/0269-8803.19.4.281.

Berberian, B., Somon, B., Sahaï, A., Gouraud, J., 2017. The out-of-the-loop Brain: a neuroergonomic approach of the human automation interaction. Annu. Rev. Contr. 44, 303–315 https://doi.org/10.1016/j.arcontrol.2017.09.010.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M.J., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imag. 18 (1), 32–42 https://doi.org/10.1109/42.750253.

Carp, J., Halenar, M.J., Quandt, L.C., Sklar, A., Compton, R.J., 2009. Perceived similarity and neural mirroring: evidence from vicarious error processing. Soc. Neurosci. 4 (1), 85–96 https://doi.org/10.1080/17470910802083167.

Chavarriaga, R., Millán, J. del R., 2010. Learning from EEG error-related potentials in noninvasive brain-computer interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. 18 (4), 381–388 https://doi.org/10.1109/TNSRE.2010.2053387.

Chavarriaga, R., Sobolewski, A., Millán, J., del, R., 2014. Errare machinale est: the use of error-related potentials in brain-machine interfaces. Front. Neurosci. 8, https://doi.org/10.3389/fnins.2014.00208.

Cracco, E., Desmet, C., Brass, M., 2015. When your error becomes my error: anterior insula activation in response to observed errors is modulated by agency. Soc. Cognit. Affect Neurosci. https://doi.org/10.1093/scan/nsv120.

de Bruijn, E.R., Schubotz, R.I., Ullsperger, M., 2007. An event-related potential study on the observation of erroneous everyday actions. Cognit. Affect Behav. Neurosci. 7 (4), 278–285 https://doi.org/10.3758/CABN.7.4.278.

de Bruijn, E.R.A., Von Rhein, D.T., 2012. Is your error my concern? An event-related potential study on own and observed error detection in cooperation and competition. Front. Neurosci. 6, https://doi.org/10.3389/fnins.2012.00008.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21 https://doi.org/10.1016/j.jneumeth.2003.10.009.

Desmet, C., Brass, M., 2015. Observing accidental and intentional unusual actions is associated with different subregions of the medial frontal cortex. Neuroimage 122, 195–202 https://doi.org/10.1016/j.neuroimage.2015.08.018.

Desmet, C., Deschrijver, E., Brass, M., 2014. How social is error observation? The neural mechanisms underlying the observation of human and machine errors. Soc. Cognit. Affect Neurosci. 9 (4), 427–435 https://doi.org/10.1093/scan/nst002.

Donchin, E., Coles, M.G.H., 1988. Is the P300 component a manifestation of context updating?. Behav. Brain Sci. 11 (03), 357 https://doi.org/10.1017/S0140525X00058027.

Donkers, F.C.L., van Boxtel, G.J.M., 2005. Mediofrontal negativities to averted gains and losses in the slot-machine task: a further investigation. J. Psychophysiol. 19 (4), 256–262 https://doi.org/10.1027/0269-8803.19.4.256.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance. Int. J. Hum. Comput. Stud. 58 (6), 697–718 https://doi.org/10.1016/S1071-5819(03)00038-7.

[4] www.grenoblecognition.fr.

Endsley, M.R., Kiris, E.O., 1995. The out-of-the-loop performance problem and level of control in automation. Hum. Factors: J. Human Fact. Ergonom. Soc. 37 (2), 381–394 https://doi.org/10.1518/001872095779064555.

Enriquez-Geppert, S., Konrad, C., Pantev, C., Huster, R.J., 2010. Conflict and inhibition differentially affect the N200/P300 complex in a combined go/nogo and stop-signal task. Neuroimage 51 (2), 877–887 https://doi.org/10.1016/j.neuroimage.2010.02.043.

Eriksen, B.A., Eriksen, C.W., 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept. Psychophys. 16 (1), 143–149 https://doi.org/10.3758/BF03203267.

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. Electroencephalogr. Clin. Neurophysiol. 78 (6), 447–455 https://doi.org/10.1016/0013-4694(91)90062-9.

Ferrez, P.W., 2007, October 25. Error-related EEG potentials in Brain-Computer Interfaces. (Génie électrique et électronique). Faculté des Sciences et Techniques de l'Ingénieur, Lausanne, Suisse.

Ferrez, P.W., Millán, J. del R., 2005. You are wrong!—automatic detection of interaction errors from brain waves. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence.

Ferrez, P.W., Millán, J. del R., 2008. Error-related EEG potentials generated during simulated brain/computer interaction. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 55 (3), 923–929 https://doi.org/10.1109/TBME.2007.908083.

Fitzgerald, K.D., Welsh, R.C., Gehring, W.J., Abelson, J.L., Himle, J.A., Liberzon, I., Taylor, S.F., 2005. Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder. Biol. Psychiatry 57 (3), 287–294 https://doi.org/10.1016/j.biopsych.2004.10.038.

Gajewski, P.D., Falkenstein, M., 2013. Effects of task complexity on ERP components in Go/Nogo tasks. Int. J. Psychophysiol. 87 (3), 273–278 https://doi.org/10.1016/j.ijpsycho.2012.08.007.

Gehring, W.J., Coles, M.G.H., Meyer, D.E., Donchin, E., 1990. The error-related negativity: an event-related brain potential accompanying errors. Psychophysiology 27 (4), S34.

Gehring, W.J., Liu, Y., Orr, J.M., Carp, J., 2011. The error-related negativity (ERN/Ne). In: https://doi.org/10.1093/oxfordhb/9780195374148.013.0120.

Gentsch, A., Ullsperger, P., Ullsperger, M., 2009. Dissociable medial frontal negativities from a common monitoring system for self- and externally caused failure of goal achievement. Neuroimage 47 (4), 2023–2030 https://doi.org/10.1016/j.neuroimage.2009.05.064.

Haatveit, B.C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., Andreassen, O.A., 2010. The validity of d prime as a working memory index: results from the "Bergen n -back task. J. Clin. Exp. Neuropsychol. 32 (8), 871–880 https://doi.org/10.1080/13803391003596421.

Holroyd, C.B., Coles, M.G.H., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109 (4), 679–709 https://doi.org/10.1037//0033-295X.109.4.679.

Jääskeläinen, I.P., Halme, H.-L., Agam, Y., Glerean, E., Lahnakoski, J.M., Sams, M., et al., 2016. Neural mechanisms supporting evaluation of others' errors in real-life like conditions. Sci. Rep. 6, 18714 https://doi.org/10.1038/srep18714.

Johnson, R., Donchin, E., 1978. On how P300 amplitude varies with the utility of the eliciting stimuli. Electroencephalogr. Clin. Neurophysiol. 44 (4), 424–437 https://doi.org/10.1016/0013-4694(78)90027-5.

Kaber, D.B., Endsley, M.R., 1997. Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. Process Saf. Prog. 16 (3), 126–131 https://doi.org/10.1002/prs.680160304.

Koban, L., Pourtois, G., Vocat, R., Vuilleumier, P., 2010. When your errors make me lose or win: event-related potentials to observed errors of cooperators and competitors. Soc. Neurosci. 5 (4), 360–374 https://doi.org/10.1080/17470911003651547.

Kok, A., 2001. On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology 38 (3), 557–577 https://doi.org/10.1017/S0048577201990559.

Kutas, M., Hillyard, S., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207 (4427), 203–205 https://doi.org/10.1126/science.7350657.

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46 (1), 50–80 https://doi.org/10.1518/hfes.46.1.50_30392.

Lewandowsky, S., Mundy, M., Tan, G., 2000. The dynamics of trust: comparing humans to automation. J. Exp. Psychol. Appl. 6 (2), 104 https://doi.org/10.11003377////1K076-898X.6.2.104.

Lopez-Calderon, J., Luck, S.J., 2014. ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front. Hum. Neurosci. 8, https://doi.org/10.3389/fnhum.2014.00213.

Madhavan, P., Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. Theor. Issues Ergon. Sci. 8 (4), 277–301 https://doi.org/10.1080/14639220500337708.

Marco-Pallarés, J., Krämer, U.M., Strehl, S., Schröder, A., Münte, T.F., 2010. When decisions of others matter to me: an electrophysiological analysis. BMC Neurosci. 11 (1), 86 https://doi.org/10.1186/1471-2202-11-86.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190 https://doi.org/10.1016/j.jneumeth.2007.03.024.

Moray, N., 1986. Monitoring behavior and supervisory control. In: In: Boff, K.R., Kaufman, L., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance, vol. 2, John Wiley & Sons, Oxford; England, pp. 1–51, Cognitive processes and performance http://psycnet.apa.org/record/1986-98619-018.

Moray, N., Inagaki, T., 2000. Attention and complacency. Theor. Issues Ergon. Sci. 1 (4), 354–365 https://doi.org/10.1080/14639220052399159.

Newman-Norlund, R.D., Ganesh, S., van Schie, H.T., de Bruijn, E.R.A., Bekkering, H., 2009. Self-identification and empathy modulate error-related brain activity during the observation of penalty shots between friend and foe. Soc. Cognit. Affect Neurosci. 4 (1), 10–22 https://doi.org/10.1093/scan/nsn028.

Ninomiya, T., Noritake, A., Ullsperger, M., Isoda, M., 2018. Performance monitoring in the medial frontal cortex and related neural networks: from monitoring self actions to understanding others' actions. Neurosci. Res. https://doi.org/10.1016/j.neures.2018.04.004.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9 (1), 97–113 https://doi.org/10.1016/0028-3932(71)90067-4.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 1–9, 2011 https://doi.org/10.1155/2011/156869.

Oostenveld, R., Praamstra, P., 2001. The five percent electrode system for high-resolution EEG and ERP measurements. Clin. Neurophysiol. 112 (4), 713–719 https://doi.org/10.1016/S1388-2457(00)00527-7.

Padrao, G., Gonzalez-Franco, M., Sanchez-Vives, M.V., Slater, M., Rodriguez-Fornells, A., 2016. Violating body movement semantics: neural signatures of self-generated and external-generated errors. Neuroimage 124, 147–156 https://doi.org/10.1016/j.neuroimage.2015.08.022.

Pavone, E.F., Tieri, G., Rizza, G., Tidoni, E., Grisoni, L., Aglioti, S.M., 2016. Embodying others in immersive virtual reality: electro-cortical signatures of monitoring the errors in the actions of an avatar seen from a first-person perspective. J. Neurosci. 36 (2), 268–279 https://doi.org/10.1523/JNEUROSCI.0494-15.2016.

Pfefferbaum, A., Ford, J.M., Weller, B.J., Kopell, B.S., 1985. ERPs to response production and inhibition. Electroencephalogr. Clin. Neurophysiol. 60 (5), 423–434 https://doi.org/10.1016/0013-4694(85)91017-X.

Riek, L.D., Rabinowitch, T.-C., Chakrabarti, B., Robinson, P., 2009. How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction. ACM, pp. 245–246 https://doi.org/10.1145/1514095.1514158.

Riesel, A., Weinberg, A., Endrass, T., Meyer, A., Hajcak, G., 2013. The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. Biol. Psychol. 93 (3), 377–385 https://doi.org/10.1016/j.biopsycho.2013.04.007.

Scheffers, M.K., Coles, M.G.H., 2000. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J. Exp. Psychol. Hum. Percept. Perform. 26 (1), 141–151 https://doi.org/10.1037//0096-1523.26.1.141.

Shane, M.S., Stevens, M.C., Harenski, C.L., Kiehl, K.A., 2009. Double dissociation between perspective-taking and empathic-concern as predictors of hemodynamic response to another's mistakes. Soc. Cognit. Affect Neurosci. 4 (2), 111–118 https://doi.org/10.1093/scan/nsn043.

Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.A., 2007. Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Hum. Factors: J. Human Fact.Ergonom. Soc. 49 (2), 227–242 https://doi.org/10.1518/001872007X312469.

Sheridan, T.B., 1992. Telerobotics, Automation, and Human Supervisory Control. MIT Press, Cambridge, Mass.

Sheridan, T.B., 1997. Eight Ultimate Challenges of Human-robot Communication. IEEE, 9–14 https://doi.org/10.1109/ROMAN.1997.646944.

Sheridan, T.B., Parasuraman, R., 2005. Human-automation interaction. Rev. Human Fact. Ergonom. 1 (1), 89–129.

Sheridan, T.B., Verplank, W.L., 1978. Human and Computer Control of Undersea Teleoperators. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Somon, B., Campagne, A., Delorme, A., Berberian, B., 2017. Performance monitoring applied to system supervision. Front. Hum. Neurosci. 11, https://doi.org/10.3389/fnhum.2017.00360.

Swets, J.A., Tanner, W.P., Birdsall, T.G., 1961. Decision processes in perception. Psychol. Rev. 68 (5), 301–340 https://doi.org/10.1037/h0040547.

Taylor, S.F., Stern, E.R., Gehring, W.J., 2007. Neural systems for error monitoring: recent findings and theoretical perspectives. Neuroscientist 13 (2), 160–172 https://doi.org/10.1177/1073858406298184.

Tomasello, M., Kruger, A.C., Ratner, H.H., 1993. Cultural learning. Behav. Brain Sci. 16 (03), 495 https://doi.org/10.1017/S0140525X0003123X.

Ullsperger, M., Fischer, A.G., Nigbur, R., Endrass, T., 2014. Neural mechanisms and temporal dynamics of performance monitoring. Trends Cognit. Sci. 18 (5), 259–267 https://doi.org/10.1016/j.tics.2014.02.009.

Van der Borght, L., Houtman, F., Burle, B., Notebaert, W., 2016. Distinguishing the influence of task difficulty on error-related ERPs using surface Laplacian transformation. Biol. Psychol. 115, 78–85 https://doi.org/10.1016/j.biopsycho.2016.01.013.

Van Schie, H.T., Mars, R.B., Coles, M.G.H., Bekkering, H., 2004. Modulation of activity in medial frontal and motor cortices during error observation. Nat. Neurosci. 7 (5), 549–554 https://doi.org/10.1038/nn1239.

Weller, L., Schwarz, K.A., Kunde, W., Pfister, R., 2018. My mistake? Enhanced error processing for commanded compared to passively observed actions. Psychophysiologye13057 https://doi.org/10.1111/psyp.13057.

Wohlschläger, A., Haggard, P., Gesierich, B., Prinz, W., 2003. The perceived onset time of self-and other-generated actions. Psychol. Sci. 14 (6), 586–591 https://doi.org/10.1046/j.0956-7976.2003.psci_1469.x.

Woods, D., Tinapple, D., 1999. W3: watching human factors watch people at work. In: Presented at the Presidential Address, Presented at the 43rd Annual Meeting of the Human Factors and Ergonomics Society., Houston, TX.