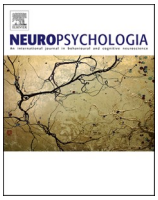
Neuropsychologia 153 (2021) 107768

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia



Predicting participants' attitudes from patterns of event-related potentials during the reading of morally relevant statements – An MVPA investigation

Manuela Hundrieser [1], André Mattes [1,*], Jutta Stahl

University of Cologne, Albertus-Magnus-Platz, 50923, Köln, Germany

ARTICLE INFO

Keywords:
Attitude
Evaluative conflict
Language processing
Moral decision
MPVA
ERP

ABSTRACT

Morality and language are hardly separable, given that morality-related aspects such as knowledge, emotions, or experiences are connected with language on different levels. One question that arises is: How rapidly do neural processes set in when processing statements that reflect moral value containing information? In the current study, participants read sentences about morally relevant statements (e.g., '*Wars are acceptable*') and expressed their (dis)agreement with the statements while their electroencephalogram (EEG) was recorded. Multivariate pattern classification (MVPA) was used during language processing to predict the individual's response. Our results show that (1) the response ('*yes*' vs. '*no*') could be predicted from 180 ms following the decision-relevant word (here *acceptable*), and (2) the attitude (pro vs. contra the topic) could be predicted from 170 ms following the topic word (here *wars*). We suggest that the successful MVPA classification is due to different brain activity patterns evoked by differences in activated mental representations (e.g. valence, arousal, etc.) depending on whether the attitude towards the topic is positive or negative and whether it is in accordance with the presented decisive word or not.

How do people come to moral decisions? This is a question that mankind has been occupied with for thousands of years. More recently, the implementation of neuropsychological methods has revived this question, as it has become possible to take a closer look at the brain processes involved. With regard to the question whether human judgements include well-considered reasoning or rather fast, intuitive gut feelings, the exact timing of decisions is of particular interest. Event-related potentials (ERPs) are useful for obtaining excellent temporal resolutions, but studies in the field of moral decision-making often make use of methods such as fMRI to localize involved brain areas, and still little is known about the timing of moral judgments (Christensen and Gomila, 2012; Wagner et al., 2017).

A prominent moral theory, one often debated in recent years, is Haidt's Social Intuitionist Model (SIM; 2001). It claims that moral judgments – individually developed values within a culture – happen intuitively, i.e., rapidly, effortlessly, and automatically. A person will only search effortfully and consciously for arguments in order to support the judgement already made; thus, according to the author, reasoning is rather a post-hoc justification process. The approach is based, amongst others, on the observations of 'moral dumbfounding', i.e., when people consider an action immoral and stick to it, even if they are incapable of finding good reasons for it (Haidt and Hersh, 2001).

Semantic priming is a phenomenon whereby the processing of a stimulus (target, e.g., *cat*) is influenced by the activated memory contents of a previous stimulus (prime, e.g., *dog*). We know that words, for example, automatically and rapidly activate associated representations founded upon prior learning, emotions, motivation, past or recent experiences, etc., so that the following related words have a processing advantage (McNamara, 2005; Neely, 1991). It is even suggested that the following words within a given context are proactively anticipated (Bar, 2011). Based on these observations, Fazio (2007) hypothesized that at least to some degree, attitudes are already represented in memory and are enabled automatically as the sum of the activated associations by processing an attitude object (e.g., *cigarettes*).

The findings mentioned above give reason to suppose that moral decisions are already influenced during the processing of language input, for example, when listening to or reading about a morally relevant issue. In other words, morally relevant words can also serve as attitude objects and automatically activate memory contents. Of course, there is also evidence for controlled decisions built on conscious, introspective

* Corresponding author. Individual Differences and Psychological Assessment, University of Cologne, Pohligstraße 1, 50969, Köln, Germany.
E-mail addresses: andre.mattes@uni-koeln.de (A. Mattes), jutta.stahl@uni-koeln.de (J. Stahl).
[1] shared first authorship.

https://doi.org/10.1016/j.neuropsychologia.2021.107768
Received 18 May 2020; Received in revised form 13 January 2021; Accepted 25 January 2021
Available online 29 January 2021

access. Nevertheless, they are more likely to happen if more time and information are available (Cunningham and Johnson, 2007; Greene et al., 2004; Schwarz, 2008). So, what if participants are asked about their attitudes towards various topics and to express their opinions for or against an issue without time pressure, but also without being given much time for introspection or elaboration? Can their judgements already be predicted from brain activity while processing the words of the sentence that they are asked to consider? Under the conditions that the processed words automatically activate related feelings, experiences, knowledge, etc., and that these activations differ depending on whether one agrees or disagrees with the issue, this should be the case. We have already outlined the first condition (e.g., priming and proactive anticipation), but there are also references in neuropsychological research to the latter of differing activation processes. For example, the contributions of left and right hemispheric structures in language processing seem to vary depending on emotional sub-processes, such as the emotional significance of the information and the emotional valence of words (for a review, see Kotz and Paulmann, 2011). A recent fMRI study of Wing et al. (2018) examined the neural mechanisms underlying subsequent memory for personal beliefs about social and political issues and found that the intensity of the ratings was linked to greater emotional arousal, with greater activity in the frontal brain regions associated with episodic memory. Moreover, the results showed brain activity differences between the response conditions, with more activity in the orbitofrontal cortex and anterior cingulate cortex for agreements compared with disagreements regarding the issues. The authors explained the results with findings that these regions are sensitive to comparisons between oneself and others, for example, on shared opinions. Studies examining ERPs also showed evidence of differentiable processes between value-consistent and value-inconsistent language processing. Leuthold, Kunkel, Mackenzie, and Filik (2015), for example, found that the reading of information associated with social-normatively acceptable versus unacceptable outputs caused a late positive potential effect (LPP; reflecting facilitated attention to emotional stimuli) with the LPP amplitude being larger for morally unacceptable than for morally acceptable sentences. The texts contained scenarios that seemed either socially appropriate (e.g., accepting the invitation of a grandfather who was terminally ill) or were considered inappropriate (e.g., accepting the invitation of a boss who makes clear advances on an employee, even though he is married, and the father of three children). Participants were instructed to simply read the texts without giving explicit judgements, therefore, the authors suspected that the scenarios triggered implicit processes of evaluative categorizations. In a further study, sentences on moral issues, such as 'I think euthanasia is an acceptable/unacceptable course of action', were presented to people who identified as strict Christians or non-Christians (van Berkum, et al, 2009). Two ERP components, the LPP and the N400 (reflecting unexpected or semantically incongruent stimuli), showed group-coherent effects following the presentation of the relevant word, which contradicted the group-specific value system. Specifically, LPP amplitudes were more positive and N400 amplitudes more negative when, for instance, strict Christians read the word "acceptable" and non-Christians read the word "inacceptable" referring to euthanasia (i.e. value-consistent statement) than when participants read value-consistent statements (i.e. "inacceptable" for Christians and "acceptable for non-Christians). The authors postulated that the effects may reflect automatic expectations regarding the personal concept of moral. Stimuli that were unexpected and contradicted the group-specific values could have led to an intensified semantic analysis or increased attention due to the emotionally aversive element of the stimuli.

In a previous study, we investigated neural correlates of moral decisions (Hundrieser and Stahl, 2016). The participants were asked to express their opinions on various moral issues. We found larger N400 amplitudes and larger LPP amplitudes for value-incongruent words compared with value-congruent words.

Based on the above-mentioned findings, the current study asked participants again to express their agreement or disagreement on morally relevant statements presented on a monitor (e.g., '*Sibling incest should be permitted*'). Unlike previous studies, we were not only interested in the critical words that were in the context of the sentence congruent or incongruent to the activated attitude (e.g., *permitted*). In addition, we were also interested in the attitude object words (e.g., *incest*), using the terminology of Fazio's attitude model (2007). Using the terminology of semantic priming, we were concerned with both the *target* words (e.g., *permitted*), hereafter referred to as decisive words, and the *prime* words (e.g., *incest*), hereafter referred to as topic words. Furthermore, our aim was to use multivariate pattern analysis (MVPA) to find the earliest time point that allowed for predicting the response outcome from the distributed spatiotemporal brain patterns while processing those two morally relevant words.

MVPA is a method that has recently been applied in cognitive neuroscience. We used an MVPA optimized for ERPs (e.g., Bode et al., 2012; Bode and Stahl, 2014). One advantage of the method is that it takes spatial and temporal aspects of neural data into account. Sentence processing is beyond finding single word meanings; it requires the simultaneous processing of visual input and its linking to contextual information. It is therefore not surprising that an extended network of brain areas is involved in a complex and dynamic interplay (Hagoort, 2017). As MVPA does not rely on a priori knowledge of specific ERP components or locations, this more data-driven explorative approach fits our purpose well (Turner et al., 2017).

1. The present study

To sum up, the aim of the present study was to identify if and when any information of morally relevant statements becomes available that predicts the response outcome, i.e., whether and how early individual judgments could be predicted from ERP signals during language processing of sentences such as 'Wars are acceptable'. Participants had to make decisions regarding various moral themes by indicating their agreement or disagreement. We expected significantly differentiable spatiotemporal brain patterns in the processing of the sentences, first, between the decisive words, depending on the post-stimulus yes or no response, and second, between the topic words, depending on the post-stimulus pro or contra attitude regarding the topic (for details see below).

2. Method

2.1. Participants

Sixty participants (M_{age} = 25.10 years, SD = 5.23; 31 females), mostly students at the University of Cologne, Germany, were paid €8.00 per hour for attendance or received course credits. All participants claimed to have good German-language skills and normal or corrected-to-normal vision. Informed consent was obtained, and the study was approved by the ethics committee of the German Psychological Society.

2.2. Stimuli and procedure

We constructed 90 sentences on nine ethical topics (war, euthanasia, genetic engineering, sibling incest, legalization of drugs, abortion, internet leaking platforms, animal experiments, and nuclear power). Each topic was used in 10 sentences with opposed sentence endings, such that five statements were composed in favour of and five against each topic (e.g., 'I think abortion is *acceptable/unacceptable*'). In each statement, a topic word that specified the attitude relevant object (here, *abortion*) was at some point followed by a decisive word (usually at the end of a sentence) which decided on the message of the statement (here, e.g., *acceptable)*. The order of sentences was randomly mixed and presented on a computer monitor word by word. Afterwards, a window appeared containing the words 'yes' and 'no' on the left and right sides

of the screen, with randomly varying sides. Participants gave their opinion on a response pad using the corresponding index finger (*yes* if they agreed with the statement, *no* if they did not). To maintain a constant and stable posture while reading and responding to the sentences, participants were seated at a distance of 50 cm in front of the monitor, with their chin on a chinrest.

The stimulus presentation was similar to the one used in one of our previous studies (Hundrieser and Stahl, 2016). Each sentence (*M* words per sentence = 6.8, range 5–10) started with the presentation of a fixation cross in the centre of the 17" VGA monitor screen (horizontal visual angle: 1.38, 1000 ms), followed by a blank screen (500 ms). The words were presented in black letters in the centre of a white screen in Arial 24-pt font. The presentation duration was 500 ms per word plus 40 ms per letter, separated by a blank screen (500 ms). The response assignment display was presented 1000 ms after the offset of the last word, with a maximum of 2000 ms to initiate the response (for illustration, see Fig. 1). Participants were instructed not to spend too much time on individual topics and to give the first response that came to their mind. Response times (RTs) were recorded using the Cedrus RB-830 response box (Cedrus Corporation).

Participants were tested individually in a separated and quiet room. Prior to testing, a standard set of written and oral instructions was given to each participant. In order to reduce socially desirable responses, the instructions emphasized the importance of honest responses and assured them that there was no objective right or wrong opinion. Furthermore, the participants were ensured that all data would be treated as strictly confidential and anonymous.

2.3. EEG data recording

EEG activity was recorded from 61 scalp electrodes positioned according to the standard international 10–20 system (Jasper, 1958; FP1, FP2, AF7, AF3, F4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C3′, C1, Cz, C2, C4, C4', C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, Poz, PO4, PO8, O1, Oz, O2; actiCAP, Brain Products). The active Ag/AgCl electrodes (actiCAP, Brain Products) were referenced against the left mastoid. Vertical and horizontal electrooculograms were placed below the right eye and 2 cm lateral to the outer canthi. The EEG was continuously recorded at a sampling rate of 500 Hz using a BrainAmp DC amplifier (Brain Products). An online band-pass filter (DC-70 Hz) was used on all channels.

EEGs were analysed offline, time-locked to the onset of the respective word presentation. The data was divided into epochs ranging from 150 ms before to 700 ms after the onset of (1) the topic words and to the onset of (2) the decisive words. The 150 ms pre-stimulus intervals were used as baselines. The length of the post-stimulus epoch (700 ms) was equal to the presentation time of the shortest topic word ("Krieg", German for "war"). All data were screened for artefacts, and contaminated trials exceeding a maximum allowed voltage step of 100 μV were rejected. The influence of eye blinks was eliminated by applying an ocular correction algorithm (Gratton et al., 1983).

2.4. Multivariate pattern classification

The goal of the approach was to distinguish between experimental conditions by means of a classifier that was trained on brain activation patterns. To this end, we used a linear support vector machine (LSVM, implemented by the LIBSVM Toolbox; Chang and Lin, 2011) classification approach that is suitable for binary-dependent variables (e.g., Bode et al., 2012; Bode and Stahl, 2014). The LSVM is a machine learning algorithm that treats each element of the EEG data for a respective analysis time window as a feature in a high-dimensional space. In machine learning, the term 'feature' is used as a general term for a set of variables or attributes (here, the corresponding EEG signal represented in a 3D matrix with the dimensions 'channels', 'time points', and 'trials'). In the first step, a classification model is generated by training the classifier on a subset of trials to determine the best highest-margin separating hyperplane (an optimal decision boundary) between the predefined classes (the two conditions). In the next step, this boundary is used to classify the remaining and unused trials producing classification accuracy (the percentage of correct classifications). In case the classification accuracy is significantly above chance, the data contains substantial information about the two conditions.

We implemented two different LSVM analyses (default regularisation parameter $C = 1$) to find the time windows that allowed for decoding the response outcome from distributed patterns of ERP data related to stimulus presentation: (1) The *decisive words* served as stimuli and provided the basis for the binary conditions belonging either to yes responses or to no responses, according to the individual's agreement or disagreement with a statement (response class-labels: *yes, no*); (2) the *topic words* were used as stimuli, belonging either to for or against judgments, according to the individual's opinion on the topic (opinion class-labels: *for, against*). Table 1 displays some exemplary variations of statement contents, participants' responses, and resulting opinions to provide a better understanding of the principles of classification.

To account for possible effects of inconsistent or uncertain opinions on an issue and to increase power, we included only topics with more than eight judgments (out of the 10 sentences) for or against a topic (e.g., nine times responding in opposition to abortion). The remaining one

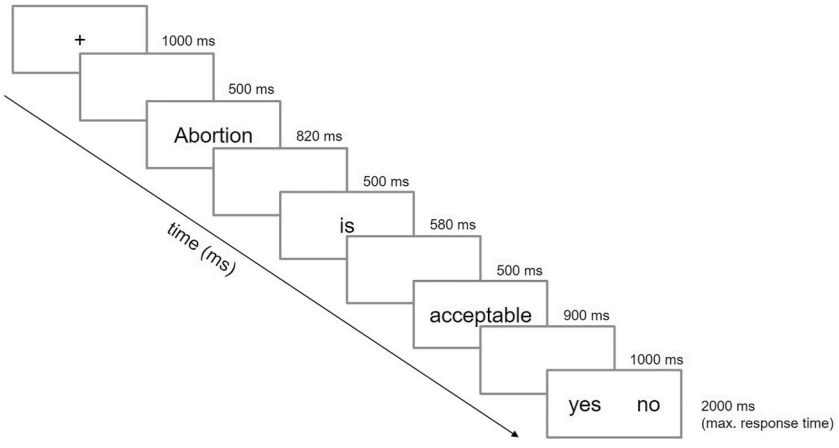


Fig. 1. Illustration of the time course of the sentence presentation. The sentences were presented word by word before the response assignment display (with randomly varying sides of 'yes' and 'no') appeared for a maximum of 2 s. The response was indicated by a key press on a response pad with the left or right index finger.

Table 1
Simplified illustration of some morally relevant sentences (including the *topic words* and the *decisive words*), with exemplary resulting opinions on the topics, depending on each statement's content in combination with the given response.

Exemplary Topic	Exemplary Statement			Response	Opinion
	Topic Word		Decisive Word		
The right to have an abortion	Abortion	is	acceptable	yes	for
	Abortion	is	unacceptable	no	for
The use of nuclear power	Nuclear power	is	useful	no	against
	Nuclear power	is	harmful	yes	against

or two trials not in line with the prevailing opinion (e.g., if the same participant responded one time in favour of the right to abortion) were excluded from the analyses. Furthermore, responses that were skipped and participants who had less than 10 trials per condition were also excluded. Due to the criteria, out of our original dataset, two participants were excluded from the first analysis ($n = 58$), and six participants were excluded from the second analysis ($n = 54$). Both analyses were based on a balanced number of trials associated with the two-class conditions. This means that we only used the smaller number of trials of the two conditions. For the condition with the larger trial numbers, trials were chosen randomly from all available trials for each participant.

For the MVPA analyses, a MATLAB version of the Decision Decoding Toolbox (DDBOX) was used (Bode et al., 2019). The moving analysis windows were applied to the brain activity patterns 150 ms before and 700 ms after the stimulus onset covering the entire epoch. The LSVM classifier analysed the data in time steps of 10 ms, with a moving window with a width of 10 ms (i.e., 150–140 ms pre-stimulus, 140–130 ms pre-stimulus, and so on, up to 690–700 ms post-stimulus). Within each time window, the data (containing five data points) from all 61 channels were transformed into vectors and served as individual features representing the spatiotemporal activity patterns. For each of the two analyses (*yes* vs. *no*; *for* vs. *against*), data from the two label types were randomly sorted into 10 sets of equal sizes. The classifier was then trained on nine sets (90% of the data), estimating the hyperplane that optimally separated exemplars from the two classes. The decision boundary was then used to classify the vectors of the remaining set (10% of the data) as either *yes* or *no*, and *for* or *against*, respectively. The percentage of trials that were classified correctly served as the resulting classification accuracy. To minimize the risk of false positive results, the entire process was first repeated using a ten-fold cross-validation procedure, so that each set was used once for testing while the remaining nine sets were used for training (Bode et al., 2014; Bode and Stahl, 2014). In addition, a conservative accuracy estimation approach was taken to rule out potential drawing biases. Therefore, the procedure was repeated 10 times, each time randomizing the pairing of trials, resulting in a total of 100 analyses. The resulting mean classification accuracy for each time window was obtained by the average of the respective 100 analyses.

In the next step, a permutation test was used for significance testing. The empirical chance distribution was obtained by repeating a similar analysis for each analysis step once again, but this time shuffling the labels randomly before classification. By adopting this approach, classification accuracy can be compared with the empirical chance distribution instead of the theoretical chance level whose limitations have lately been criticised (Combrisson and Jerbi, 2015; Turner et al., 2017).

First, analyses were performed on each participant's EEG data separately. Then, statistical analyses were performed at the group level. For each 10-ms time window, the results (classification accuracy versus empirical chance, composed of the average accuracies of the participant level) were tested with paired sample *t*-tests. Corrections for multiple comparisons were performed with cluster-based permutation tests (global alpha level of 0.05; Bullmore et al., 1999; Maris and Oostenveld, 2007). Furthermore, for each time window in which the labels were successfully classified, the individual absolute feature weights (*z*-standardised) were extracted and assigned to each of the 61 channels. Note that these values allow a rough estimation of their importance for the classification but cannot be interpreted as the sources of this information (Haufe et al., 2014; Turner et al., 2017). For example, it is possible that significant channels are only important for the classification because they reduce the noise of irrelevant variance for the features that carry the relevant information (see Bode and Stahl, 2014).

3. Results

3.1. Word length

The effects of word length were examined to see whether the various words classified as *yes* or *no* and *for* or *against* differ in word length. We used two different measures of word length: length in letters and length in phonemes, i.e., the smallest sound units in language (Bijeljac-Babic et al., 2004). The descriptive statistics are given in Table 2.

For the two different word types, separate ANOVAs were performed for the word-length variables. The ANOVAs for the decisive words showed neither significant differences between the *yes* and *no* word lengths in letters, $F(1, 57) = 3.90$, $p = .053$, nor in phonemes, $F(1, 57) = 2.01$, $p = .162$. However, the ANOVAs for the topic words showed significant length effects between the *for* and *against* conditions, both in letters, $F(1, 53) = 170.10$, $p < .001$, $\eta^2 = 0.76$, and in phonemes, $F(1, 53) = 190.50$, $p < .001$, $\eta^2 = 0.78$. This means that the topic words with an agreeing attitude were on average 1.6 letters and 1.5 phonemes longer than the words with a disagreeing attitude (see Table 2).

3.2. Behavioral data

First, we investigated whether the different responses classified as *yes* or *no* and, furthermore, classified as *for* or *against* differed with respect to response rate (i.e., how often a certain response was made) and to reaction times (see also Table 3).

The respective ANOVAs for response rate showed neither significant differences between the number of agreements and disagreements regarding the statements (*yes* vs. *no*), $F(1, 57) = 2.97$, $p = .090$, nor between the number of pro or contra opinions regarding the statement topics (*for* vs. *against*), $F(1, 53) = 0.02$, $p = .896$. Furthermore, the ANOVAs for reaction times showed neither significant differences between the speed of *yes* and *no* responses, $F(1, 57) = 0.83$, $p = .367$, nor between the response speed of *for* and *against* opinions, $F(1, 53) = 2.27$, $p = .138$.

3.3. Multivariate pattern classification

Decisive words. A linear SVM classifier was used to predict participants' post-stimulus agreement or disagreement (*yes* or *no*) from the

Table 2
Descriptive statistics for the different word lengths in letters and phonemes: first, for the *yes* and *no* responses according to the agreement or disagreement with the sentence (n = 58); and second, for the *for* and *against* opinions according to the judgment regarding the sentence topic (n = 54). The decisive words were mostly adjectives and the topic words mostly nouns.

	Letters	*S.E.M.*	Phonemes	*S.E.M.*
Decisive Word				
"yes"	8.6	0.7	8.1	0.7
"no"	8.5	0.9	8.2	0.8
Topic Word				
"for"	11.4	2.2	10.3	2.0
"against"	9.8	2.9	8.8	2.5

Note. S.E.M. = standard error of the mean.

Table 3
Response rate and mean reaction time results for the different classification types: first, for the *yes* and *no* responses according to the agreement or disagreement with the sentence (n = 58); and second, for the *for* and *against* opinions according to the judgment regarding the sentence topic (n = 54).

	RR	*S.E.M.*	RT (ms)	*S.E.M.*
Response				
"yes"	29	0.96	659	18
"no"	28	0.95	667	18
Opinion				
"for"	31	1.61	675	20
"against"	31	1.65	662	19

Note. RR = response rate; RT = reaction time; S.E.M. = standard error of the mean.

spatiotemporal patterns of their EEG data, according to the statements' decisive words (e.g., 'I think abortion is *acceptable*'). Two statistically significant clusters were found between 180–260 ms and 470–490 ms following stimulus presentation. The trained LSVM performed significantly better than chance in these time-windows in distinguishing the *yes* responses from the *no* responses (see Fig. 2).

Topic words. Next, we used the classifier to predict participants' *for* or *against* opinions about the morally relevant topics. To this end, the EEG data were time-locked to the statements' topic word (e.g., 'I think *abortion* is acceptable'). Participants' pro or contra judgments of the topic of the statement could be predicted significantly above chance between 170 and 360 ms after stimulus presentation. In this time window, four statistically significant clusters were found: 170–200 ms, 220–240 ms, 260–280 ms, and 330–360 ms (see Fig. 3).

In sum, we found that during processing of the sentences, EEG signals

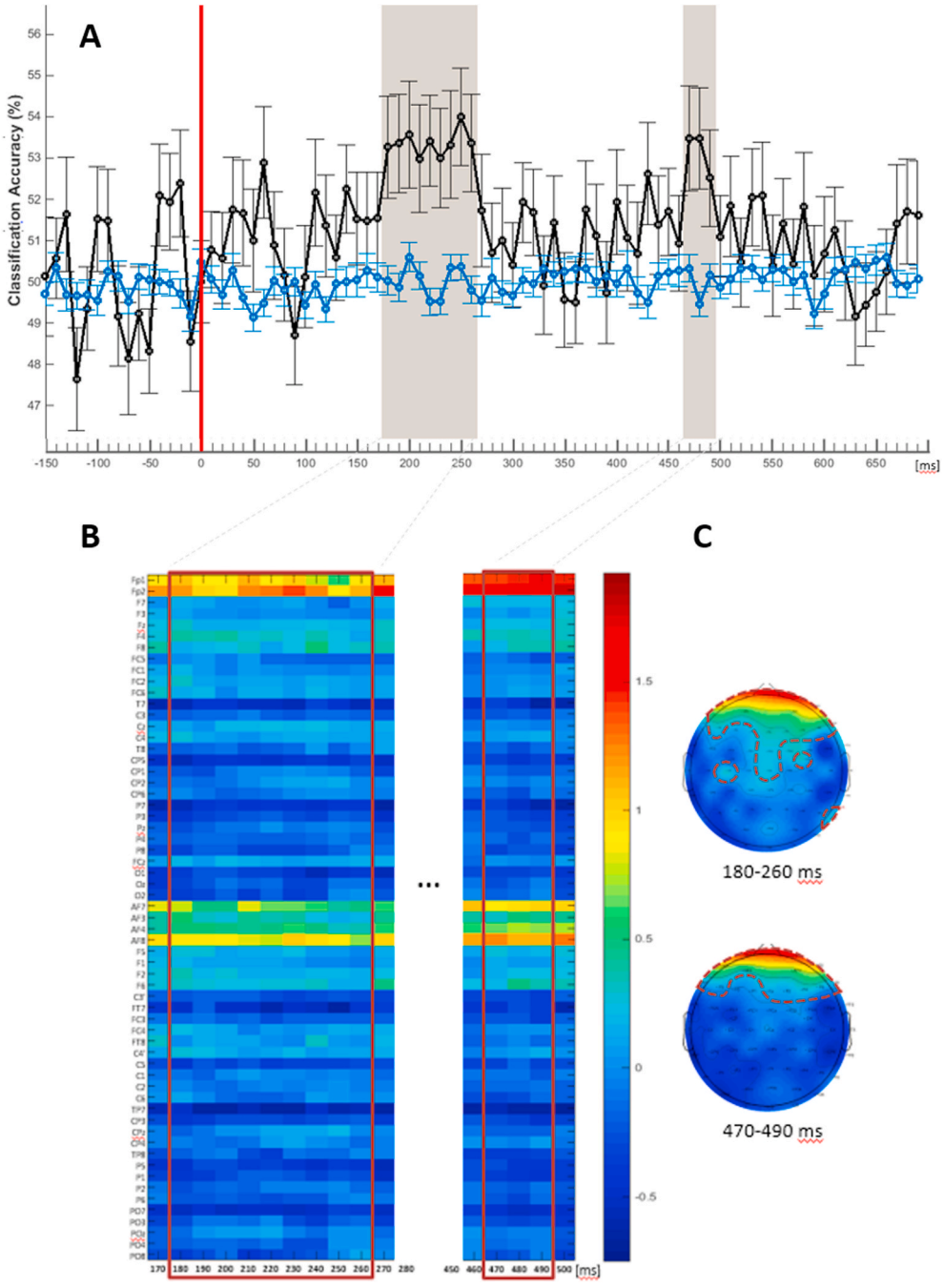


Fig. 2. Spatiotemporal decoding of the statements' decisive words (e.g., *acceptable*). (A) LSVM classification was used to predict participants' responses (*yes* vs. *no*) from distributed patterns of their ERPs while processing the morally relevant statements. The black line represents the classification accuracy, the blue line the empirical chance distribution. The grey bars denote the significant differences between the two distributions (180–260 ms, 470–490 ms; N=58). (B) Z-standardised, averaged absolute feature weights for the two significant time-windows. Red and yellow channels indicate high absolute feature weights; light blue and dark blue channels indicate low absolute feature weights. (C) Scalp maps of the z-standardised absolute feature weights averaged across the significant time-steps. Predictive channels reaching significance (Bonferroni-corrected) are indicated by red dotted frames. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

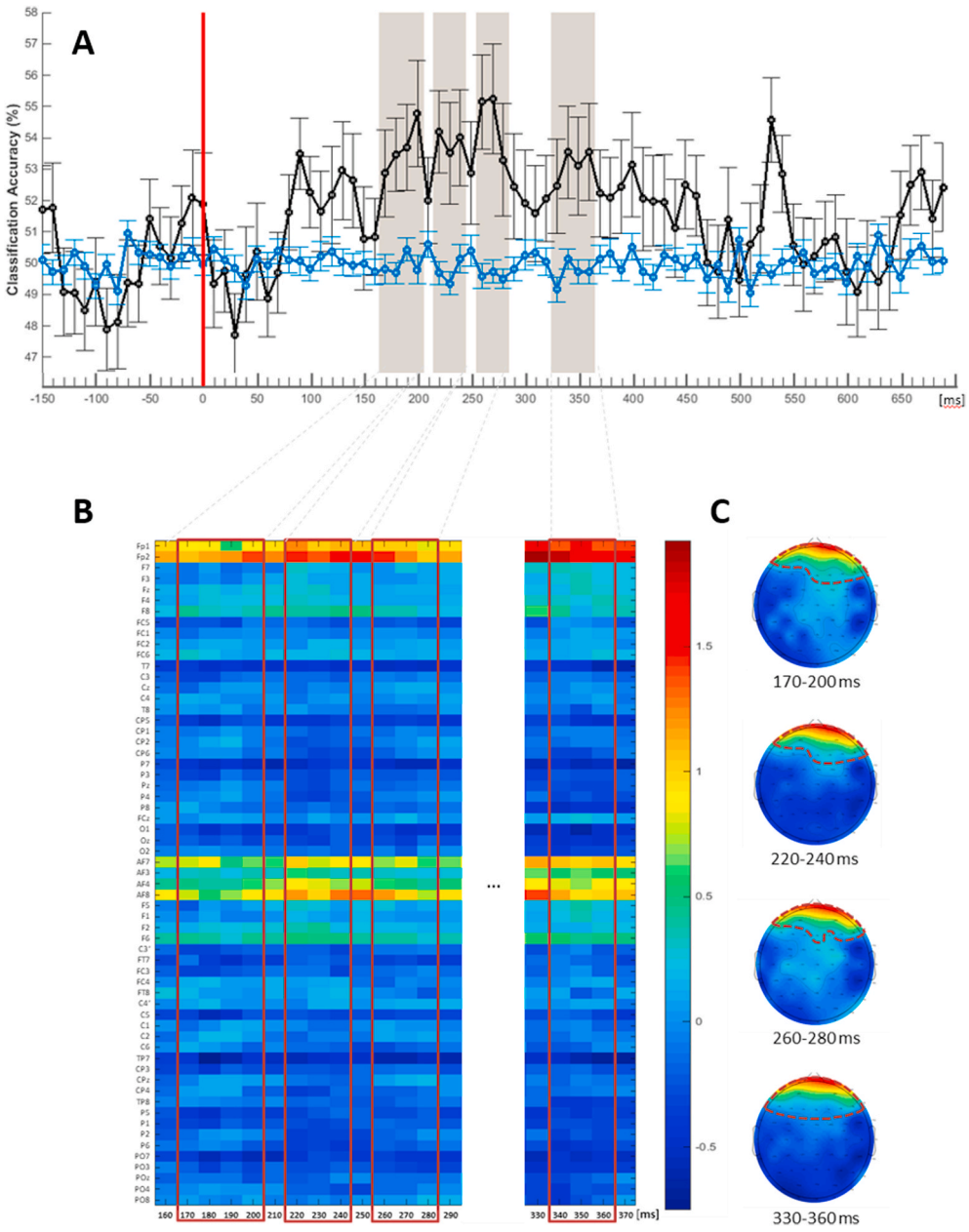


Fig. 3. Spatiotemporal decoding of the statements' topic words (e.g., *abortion*). (A) LSVM classification was used to predict participants' opinion (*for* or *against*) from distributed patterns of their ERPs while processing the morally relevant statement. The black line represents the classification accuracy, the blue line the empirical chance distribution. The grey bars denote the significant differences between the two distributions (170–200 ms, 220–240 ms, 260–280 ms, 330–360 ms; N=54). (B) Z-standardised, averaged absolute feature weights for the four significant time-windows. Red and yellow channels indicate high absolute feature weights; light blue and dark blue channels indicate low absolute feature weights. (C) Scalp maps of the z-standardised absolute feature weights averaged across the significant time-steps. Predictive channels reaching significance (corrected) are indicated by red dotted frames. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

of both the topic words and the statements' decisive words carried information that predicted participants' post-stimulus responses.

3.4. Feature weights and post-hoc ERP analyses

We also analysed the feature weights for the significant classification time windows. The spatial distributions of the *z*-standardised absolute feature weights for the significant time windows are presented in Figs. 2B and 3B. Visually, there are many similarities between the absolute feature weight maps for all of the statistically significant clusters of the decisive word and the topic word classifications (see Figs. 2B and 3B). Significant Bonferroni-corrected ($p < .05$) channels were found mostly over prefrontal cortices, with electrodes Fp1 and Fp2 reaching the highest absolute feature weights in all conditions (see also Figs. 2C and 3C; high values are indicated in red colours). As already mentioned above, the values cannot be interpreted as the underlying neural sources, therefore, their interpretation is merely hypothetical. Since, however, the feature weights correspond to the relative classifying contributions, they can be an indication of the importance of the channels. Therefore, to gain further insight into the processes behind these results, we investigated the ERPs at Fp1 and Fp2, averaged across the electrodes and time points within each significant MPVA classification time window in a series of *t*-tests. Averaged event-related potentials time-locked to the stimulus onset are shown in Fig. 4.

Decisive words. The mean amplitude in the first significant MVPA classification time window between 180 and 260 ms showed significant differences between the trials with *yes* (M ± SEM: 0.71 ± 0.79 μV) versus *no* responses (2.66 ± 0.95 μV), $t(57) = -2.66$, $p = .010$, $d = .349$. This means that ERP amplitudes were more positive for decisive word stimuli with following no responses than with following yes responses. There was no significant difference between the two conditions for the ERP amplitudes during the second time window from 470 to 490 ms, $t(57) = -0.18$, $p = .856$ (*yes* $M = 5.53$ μV ± 1.28, *no* $M = 5.76$ μV ± 1.22).

Topic words. The *t*-tests for the topic words in the four selected time windows, which were found to be significant in the MVPA analyses, revealed no significant differences between *for* and *against* conditions

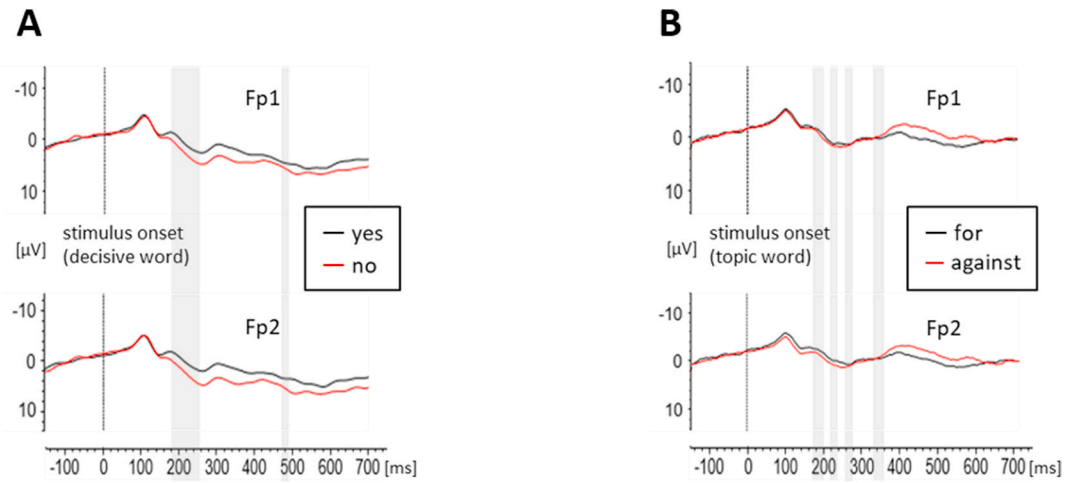


Fig. 4. Grand-average, event-related potentials time-locked to the onset of (A) the decisive words, separated for *yes* versus *no* responses; and (B) the topic words, separated between *for* and *against* opinions, illustrated at the channels with the highest feature weights (Fp1, Fp2). Note that for illustration purposes, we display the waveforms for the Fp1 and Fp2 separately, but in the analysis of the significant MVPA time windows, we used the average of both electrodes.

between 170 and 200 ms, $t(53) = 1.78$, $p = .081$ (*for* 0.09 μV ± 0.92, *against* −1.36 μV ± 0.83); between 220 and 240 ms, $t(53) = 1.38$, $p = .174$ (*for* 1.99 μV ± 0.91, *against* 0.65 μV ± 1.03); between 260 and 280 ms, $t(53) = 1.51$, $p = .138$ (*for* 1.45 μV ± 1.06, *against* 0.08 μV ± 1.14); and between 330 and 360 ms, $t(53) = 1.13$, $p = .265$ (*for* -0.18 μV ± 1.25, *against* −1.23 μV ± 1.02).

4. Discussion

In this study, using MVPA, we examined EEG patterns for two different stimulus types while participants were reading morally relevant sentences: first, words that completed the content of a sentence to predict the participant's yes or no response, and second, words that indexed the sentence topic to predict the participant's for or against attitude towards the issue. We found that subsequent decisions on moral topics could be predicted successfully from distributed patterns of brain activity recorded during language processing from less than 200 ms after stimulus presentation. This early time point is remarkable since research in language processing shows that early ERP components up to 200 ms are found, for example, for perceptual or syntax-related processing (Hagoort, 2017; Hillyard et al., 1998; Vogel and Luck, 2000). Word recognition, e.g., identifying a particular word from other representations in long-term memory, takes about 150 ms at the earliest up to 300 ms or even later (e.g., Klimovich-Gray et al., 2019; Pylkkänen and Marantz, 2003).

4.1. Processing of the decisive words

Predictions of *yes* and *no* responses were possible from 180 to 260 ms, as well as around 480 ms after stimulus onset (see Fig. 2A). In a previous study, we found, in accordance with our hypotheses, an N400-effect (300–500 ms) and an LPP-effect (500–800 ms) for value-incongruent words compared with value-congruent words (comparable to the decisive words in this study; Hundrieser and Stahl, 2016). These results seem to be confirmed by the second significant time window in our present findings, obtained with a different method in a new sample.

Furthermore, the highest absolute feature weights in the present study were found at frontal channels, and our post-hoc ERP-analyses for the frontal channels showed a significant difference between the *yes* and *no* conditions in the early significant time window (see Fig. 4). This finding matches the P2 component, a positive potential, often peaking around 200 ms and distributed around the frontal region of the scalp. The P2 has been found to be modulated by attention, language context information, and memory processes, comparing, for example, incoming stimuli with stored memory (Luck and Hillyard, 1994). Leuthold et al. (2015), for instance, found pronounced P2 amplitudes in response to socio-normative violations, suggesting attentional allocation for unexpected stimuli (see also Lu et al., 2019). In our study, the decisive words appeared in the course of a sentence where the moral issue was already named by the topic words. Assuming that the topic words primed associated memory contents, leading to an individual attitude, the processing of a semantically activated target word should be facilitated (Cunningham and Johnson, 2007; Fazio, 2007). Thus, reading a sentence that ends in conflict with the activated attitude could attract more attention, which might be reflected in a P2 effect for the decisive words.

Another possible explanation for the differences we found between the decisive words may not only be due to the ease of word processing but also to different induced feelings when processing a sentence that contradicts or corresponds to the participant's own opinion. For example, Carretié et al. (2001) found that the P200 amplitude increased as the stimulus was evaluated more negative. Perhaps, disagreeing with the statement in our study evoked a more pronounced P200 in response to the decisive word than agreeing with the statement, because disagreement is associated with negative affect. Further support for this interpretation is provided by Wing et al. (2018). The authors found differences in individuals' brain activity between agreements and disagreements on social and political issues. Participants showed more activity in the orbitofrontal and anterior cingulate cortex when they read about political opinions they agreed with (Gozzi et al., 2010). In communication research, it has been shown that achieving an agreement about attitudes and beliefs is perceived favourably, whereas disagreements are experienced negatively (Michaels et al., 2013). Unpleasantness experienced during decision-making on moral dilemmas has been shown to correlate with the P260 amplitude over fronto-polar and frontal locations (Sarlo et al., 2012), suggesting that the P2 effect that we found at comparable electrode sites and in comparable time windows might reflect the emotional state of the participants during decision-making.

4.2. Processing of the topic words

A matter of particular interest in our study was the investigation of the topic words. The MVPA classifier was able to predict the participants' *for* or *against* attitudes successfully 170 ms after stimulus onset (see Fig. 3A). Again, the significant predictions were accompanied by high feature weights for channels over the prefrontal cortex. Interestingly, the ERP amplitudes revealed no significant differences between

the conditions. The missing sensitivity of P2 seems to make sense as different word analyses take place during sentence processing. In the course of each of the 90 randomly presented sentences, the topic word appeared to the participant mostly as the first clue on the topic, such that it is very unlikely that morally related processes specific to this topic occurred before the presentation of the word. Therefore, the mere presentation of the topic word should not have elicited any type of stimulus (un)expectancy, making the emergence of a P2 effect equally unlikely (Lu et al., 2019; Luck and Hillyard, 1994). Given that each topic word was presented multiple times in the course of the experiment, it is valid to assume that participants formed some sort of expectation regarding which topics might occur in the course of the experiment in general. However, considering the random order of topic word presentation, it is highly unlikely that this produced the observed MVPA results.

Nevertheless, using MVPA, there was evidence that the attitudinal tendency *for* or *against* an issue could be predicted from brain patterns during topic word processing. One possible explanation could be the differential emotional and evaluative content of the stimuli. In a review article on neural correlates of emotional word processing, Citron (2012) describes two main findings: First, there are indications that the emotional intensity (i.e., arousal) of stimuli triggers rapid, unconscious processes, indicated by amygdala activation and early ERP components. Second, the valence, i.e., the positive or negative evaluative content of stimuli, occurs in a later conscious process, indicated by later, more controlled activations of the orbitofrontal and prefrontal cortex and paired with later ERP components. Results from an affective word list study showed a U-shaped function between emotional arousal and emotional valence of words, with higher arousal for emotional words compared to neutral ones, and, in addition, higher arousal ratings for negative than positive words (Võ et al., 2009). Negative stimuli are often associated with processes related to attention orientation and risk prediction. Positive stimuli, on the other hand, are associated with positive affect and reward (Citron, 2012).

Although our study did not assess ratings, it is assumed that even with the moral themes one agrees with (such as being in favour of *abortion*), the concepts themselves would not actually have been rated 'positively', in contrast to words such as *love* or *happiness* in the studies conducted in terms of word ratings (e.g., Garcia et al., 2012; Rohr and Rahman, 2015). Note that this does not mean that the concepts we used can be considered neutral or less emotional. Instead, it is more likely that they can be categorized as ambivalent, since they are cognitively complex, with multiple factors and potentially opposing perspectives activating both, positive and negative evaluation components (Cunningham et al., 2003, 2004; Kaplan, 1972; Nohlen et al., 2014). Cunningham et al. (2004) for example asked participants to make evaluative judgments (good vs. bad) about socially relevant positive, negative, or ambivalent concepts (e.g., *welfare*, *murder*, or *abortion,* respectively). They found that stimuli rated as ambivalent were indicated by high emotional intensity, and correlated with activation brain areas such as the anterior cingulate and prefrontal cortex which are associated with cognitive control. The authors suggest that these brain areas were activated because they were involved in the attempt to solve the evaluatively inconsistent/conflicting information (see also Cunningham et al., 2003; Nohlen et al., 2014).

As stated before, we assume we have used rather ambivalent stimuli. In the analysis, however, only topics on which participants expressed relatively consistent opinions across the 10 sentences were included, with the valence in either the pro or contra direction prevailing. In the case of pro opinions, the positive aspects of the concepts should predominate, and in the case of contra opinions, the negative ones should predominate. Furthermore, participants had on average an equal number of *for* and *against* attitudes towards the topics they were presented with (see Table 3). As discussed above, previous work has shown that negative attitudes come along with a higher level of arousal than positive attitudes (e.g. Schmidtke et al., 2014; Võ et al., 2009). Accordingly, it should be possible to predict the participants' attitude not only from valence (positive valence – *for* attitude; negative valence – *against* attitude), but also from arousal. Higher arousal suggests that the participant would indicate a response *against* the topic while lower arousal suggests that the participant would indicate a response *for* the topic. As previous work has shown that different levels of arousal produce different patterns of brain activity (e.g. Canli et al., 2000; Cunningham et al., 2004), we suggest that this might be what allowed our classifier to decode the participant's attitude from the EEG signal. The same applies to valence (Cunningham et al., 2004; Nohlen et al., 2014). Considering that multiple studies propose that arousal is processed faster than valence (e.g. Citron, 2012), the earlier time window in which the classifier successfully decoded participants' attitude might reflect differences in arousal, while the later time window could reflect differences in valence between the for and against attitudes (see Citron, 2012).

4.3. Limitations

As a limitation, it should be noted that there is evidence of dissociations between implicit and explicit measures in attitude research. Therefore, one could question the design of the study per se, since the MVPA classifications were related to the actual (and maybe socially desirable) responses, but they may not correspond to the individual's implicit attitude. However, the gap between implicitly and explicitly measured results depends on the task, and it is lower if participants are asked to decide on the basis of their 'gut feeling' (see Gawronski and Bodenhausen, 2012). In the present study we have placed great emphasis on assuring participants not to worry about their responses, since there were no objectively right or wrong decisions. This proceeding may also have contributed to rather intuitive and less socially desirable answers.

Furthermore, it should be mentioned that a difference in the word lengths of the topic words could be found, so that a significant difference of about 1.5 letters on average occurred between the *for* and the *against* conditions. Word length affects, for example, the number of eye fixations, or the word-identification time, and could therefore influence language processing. However, in studies investigating word-length differences, stimuli are often distinguished in so-called short words, which usually contain about three to four letters, and long words, which contain about six to eight letters (Bertram and Hyönä, 2002; Bijeljac--Babic et al., 2004). Thus, in our study, the words in both conditions can be classified as rather long (9–11 words). In addition, Bijeljac-Babic et al, (2004) were able to show that the length effects in visual word recognition are almost eliminated in adult ages compared, for example, to children. As our sample consisted of adults, and the actual difference between the conditions seems rather small, it is likely to be of no relevance. Furthermore, we only investigated effects that were observed before the offset of the shortest word ("Krieg"; German for "war"), i.e. in the interval from topic word onset to 700 ms after the onset. Accordingly, the differences in word length cannot explain these very early effects.

Parts of our suggestions rely on the observation that arousal is higher for negative than for positive stimuli. Unfortunately, we did not obtain valence and arousal ratings in our sample and have to refer to findings in the literature. However, this observation is very consistent in German affective word databases and has been reported in many studies (e.g. Schmidtke et al., 2014; Võ et al., 2009). Future studies might collect valence and arousal ratings and use MVPA to predict these on a trial-by-trial basis from the brain activity patterns in response to topic words. This would allow to disentangle whether the successful decoding of the *for* or *against* attitude from the brain activity following the topic word was really due to differences in valence and/or arousal associated with the topic word.

5. Conclusion

Our MVPA findings provide evidence of substantial brain pattern

associations according to the later-made responses to both types of words (topic words and decisive words). Fig. 5 summarizes our results and discussion.

We suggest that while reading a socially or morally relevant sentence, such as 'Wars are acceptable', the topic-relevant word (or attitude object; here, *war*) might activate mental associations (e.g., *war crimes*, *civilian casualties*) which are stored in memory. These memory contents can be based on more or less important, motivating, neutral, positive or negative information, emotion, experiences, and so on. Depending on the content of these components, arousal and valence arise (e.g., *high* arousal, with a *negative* valence), and an individual attitude emerges as the summary evaluation (e.g., *opponent of war*). We suppose that depending on whether the opinion is *for* or *against* the topic, different patterns in arousal and/or valence emerge which can be decoded from the brain activity patterns.

Recently, models of language processing have referred to feedforward connections and predictive coding, assuming that the brain is generating proactively probabilistic predictions about upcoming information affected by prior context (Bar, 2011; Hagoort, 2017; Klimovich-Gray et al., 2019). In most studies, the prediction refers to the semantic content of a sentence context, i.e., starting words of a sentence lead to corresponding expectations concerning the end of it (Kutas and Federmeier, 2011; Nieuwland and Van Berkum, 2006). In our study, a proactive prediction seems very reasonable for the decisive words, as they were presented in the course of the sentence at some point after the topic words. But it seems also quite reasonable that at least in the course of the experiment, both directions of sentence endings were similarly predictable, since the presentation of the 90 sentences was randomized and ended in one direction or the other (i.e., 'Wars are *acceptable*' or 'Wars are *unacceptable*'). Nevertheless, the motivational relevance of the task was to make decisions based on one's own opinion. Therefore, it is more reasonable to assume that 'prediction' in this context contributes to a preparatory comparison between the own activated attitude ('I am an opponent of war') and the incoming stimulus (e.g., *acceptable*). A sentence ending in accordance with the attitude can be processed more easily and even be perceived favourably, ending in a *yes* response. A mismatch between these two, reflected also in an enlarged P2 amplitude, could be indicative of a *no*-response preparation. Although no actual motor preparation was possible at the time of the sentence word presentation (because the response assignment display was presented 1000 ms after the offset of the last word and the sides of the words *yes* and *no* on the screen varied randomly), the cognitive decision for agreement or disagreement was at least possible by presenting the last critical word of the sentence.

There are multiple theoretical frameworks that explain moral judgement, e.g. the universal moral grammar theory (Hauser, 2006; Mikhail, 2007, 2011) or the dual-process and single-process theories (Greene et al., 2001 and Moll et al., 2008, respectively). Our results may be reconciled with the single-process and dual-process theories of moral judgement. The dual-process theory states that moral judgements are the result of two separate systems that compete with each other: a fast and rather uncontrolled system producing an affective response (limbic system) and a slower and more controlled system producing a cognitive response (prefrontal cortex and parietal areas) to a moral dilemma (Greene et al., 2004; Greene et al., 2001). In contrast, the single-process theory postulates that cognitive and affective aspects of a moral

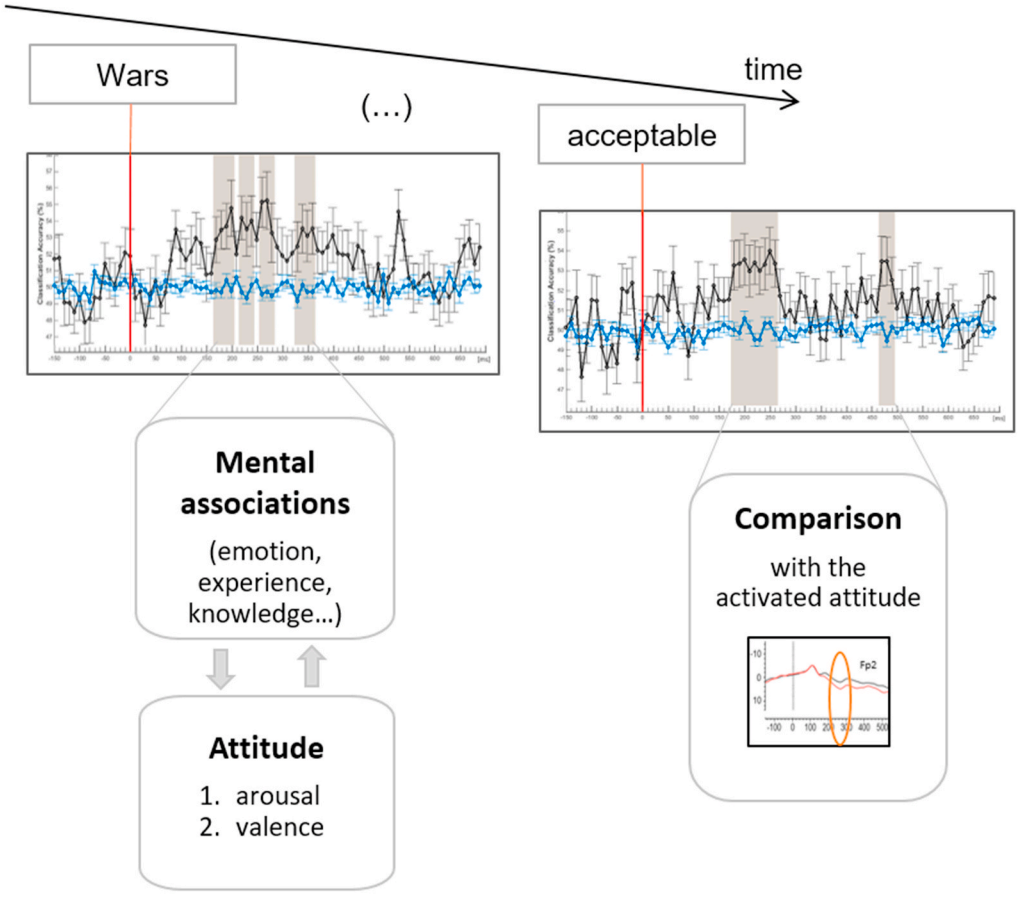


Fig. 5. Illustration of the study results and the assumed associated language processes.

judgement may not be isolated from each other, but go along with each other and form a compound (Moll et al., 2008; Moll et al., 2005). The competition is here not among an affective and a cognitive response, but among different cognitive-affective options. Our paradigm was designed to investigate the temporal dynamics of moral judgement processing, thus, it will not fully allow identifying the source of the moral conflict (affective vs. cognitive response; among different cognitive-affective options). However, it might give some first ideas, about some time-course related aspects. The classifier was able to predict the participants *for* or *against* attitude roughly 200 ms after the topic word onset. We suppose that the processes in this time window are mainly affective given that the single word without any further context may have been evaluated more generally (e.g. in terms of arousal and/or valence). At this point the participants did not know the end of the sentence, thus, no decision has to be made, which would require more cognitive resources. To further corroborate this notion, a future study might collect valence and arousal ratings and train a classifier to predict these ratings from the brain activity patterns following the topic word onset (see also Limitations). In addition, the classifier successfully decoded the participants' 'yes' and 'no' response starting roughly 200 ms following the presentation of the decisive word. In this time interval, both affective and cognitive processing might occur: Cognitive processes may involve a preparatory comparison between the own activated attitude and the incoming stimulus, i.e. the decisive word. Affective processes were presumably activated, because depending on the result of this comparison, negative/positive affect may emerge. Although our study was not designed to disentangle the dual-process and single-process theories, it contributes to understanding *when* processes occur that are related to moral judgement.

To sum up, the use of a multivariate technique allowed us to detect systematic differences in distributed brain activity patterns and found evidence that moral judgments can be predicted during very early reading processes, even long before the actual judgement is delivered.

Credit author statement

MH and JS conceptualised the study. MH scripted the paradigm, collected and analysed the data and visualised the results. AM contributed to analysis of the data and visualization of the results. MH, JS and AM interpreted the data. MH wrote the first draft of the manuscript. AM assisted in writing the first draft. JS and AM edited the first draft. AM was in charge of the revision and changed the manuscript according to the reviewers' suggestions. MH and JS edited the revised manuscript. MH was responsible for project administration. JS acted as supervisor.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of competing interest

None.

References

Bar, M., 2011. Predictions in the Brain: Using Our Past to Generate a Future. Oxford University Press, New York.

Bertram, R., Hyönä, J., 2002. The length of a complex word modifies the role of morphological structure: evidence from eye movements when reading short and long Finnish compounds. J. Mem. Lang. 48, 615–634. https://doi.org/10.1016/S0749-596X(02)00539-9.

Bijeljac-Babic, R., Millogo, V., Farioli, F., Grainger, J., 2004. A developmental investigation of word length effects in reading using a new on-line word identification paradigm. Read. Writ. 17, 411–431.

Bode, S., Bennett, D., Stahl, J., Murawski, C., 2014. Distributed patterns of event-related potentials predict subsequent ratings of abstract stimulus attributes. PloS One 9 (10), e109070.

Bode, S., Feuerriegel, D., Bennett, D., Alday, P.M., 2019. The decision decoding ToolBOX (DDTBOX) - a multivariate pattern analysis Toolbox for event-related potentials. Neuroinformatics 17 (1), 27–42. https://doi.org/10.1007/s12021-018-9375-z.

Bode, S., Sewell, D.K., Lilburn, S., Forte, J.D., Smith, P.L., Stahl, J., 2012. Predicting perceptual decision biases from early brain activity. J. Neurosci. 32 (36), 12488–12498.

Bode, S., Stahl, J., 2014. Predicting errors from patterns of event-related potentials preceding an overt response. Biol. Psychol. 103, 357–369.

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.

Canli, T., Zhao, Z., Brewer, J., Gabrieli, J.D.E., Cahill, L., 2000. Event-related activation in the human amygdala associates with later memory for individual emotional experience. J. Neurosci. 20 (19), RC99, 1-5.

Carretié, L., Mercado, F., Tapia, M., Hinojosa, J.A., 2001. Emotion, attention, and the 'negativity bias', studied through event-related potentials. Int. J. Psychophysiol. 41 (1), 75–85.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3), 27.

Christensen, J.F., Gomila, A., 2012. Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neurosci. Biobehav. Rev. 36, 1249–1264.

Citron, F.M.M., 2012. Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. Brain Lang. 122, 211–226. https://doi.org/10.1016/j.bandl.2011.12.007.

Combrisson, E., Jerbi, K., 2015. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods 250, 126–136. https://doi.org/10.1016/jneumeth.2015.01.010.

Cunningham, W.A., Johnson, M.K., 2007. Attitudes and evaluation: toward a component process framework. In: Harmon-Jones, E., Winkielman, P. (Eds.), Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior. Guilford Press, New York, pp. 227–245.

Cunningham, W., Johnson, M., Gatenby, C., Gore, J., Banaji, M., 2003. Neural components of social evaluation. J. Pers. Soc. Psychol. 85, 639–649. https://doi.org/10.1037/0022-3514.85.4.639.

Cunningham, W.A., Raye, C.L., Johnson, M.K., 2004. Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. J. Cognit. Neurosci. 16, 1717–1729. https://doi.org/10.1162/0898929042947919.

Fazio, R.H., 2007. Attitudes as object-evaluation associations of varying strength. Soc. Cognit. 25, 603–637.

Garcia, D., Garas, A., Schweitzer, F., 2012. Positive words carry less information than negative words. EPJ Data Science 1 (3), 1–12. https://doi.org/10.1140/epjds3.

Gawronski, B., Bodenhausen, G.V., 2012. Self-insight from a dual-process perspective. In: Vazire, S., Wilson, T.D. (Eds.), Handbook of Self-Knowledge. The Guilford Press, New York, pp. 22–38.

Gozzi, M., Zamboni, G., Krueger, F., Grafman, J., 2010. Interest in politics modulates neural activity in the amygdala and ventral striatum. Hum. Brain Mapp. 31 (11), 1763–1771. https://doi.org/10.1002/hbm.20976.

Gratton, G., Coles, M., Donchin, E., 1983. A new method for off-line removal of ocular artifact. Electroencephalogr. Clin. Neurophysiol. 55, 468–484. https://doi.org/10.1016/0013-4694(83)90135-9.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgement. Neuron 44 (2), 389–400.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, T., Cohen, J.D., 2001. An fMRI investigation of emotional engagement in moral judgement. Science 239 (5537), 2105–2108.

Hagoort, P., 2017. The core and beyond in the language-ready brain. Neurosci. Biobehav. Rev. 81, 194–204. https://doi.org/10.1016/j.neubiorev.2017.01.048.

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. 108 (4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814.

Haidt, J., Hersh, M., 2001. Sexual morality: the cultures and emotions of conservatives and liberals. J. Appl. Soc. Psychol. 31, 191–221. https://doi.org/10.1111/j.1559-1816.2001.tb02489.x.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.

Hauser, M.D., 2006. Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong. HarperCollins, New York.

Hillyard, S.A., Vogel, E.K., Luck, S.J., 1998. Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philosophical Transactions of the Royal Society: Biol. Sci. 353, 1257–1270.

Hundrieser, M., Stahl, J., 2016. How attitude strength and information influence moral decision making: evidence from event-related potentials. Psychophysiology 53, 678–688. https://doi.org/10.1111/psyp.12599.

Jasper, H.H., 1958. Report of the committee on methods of clinical examination in electroencephalography. Electroencephalogr. Clin. Neurophysiol. 86, 176–182.

Kaplan, K.J., 1972. On the ambivalence-indifference problem in attitude theory and measurement: a suggested modification of the semantic differential technique. Psychol. Bull. 77 (5), 361–372.

Klimovich-Gray, A., Tyler, L.K., Randall, B., Kocagoncu, E., Devereux, B., Marslen-Wilson, W.D., 2019. Balancing prediction and sensory input in speech comprehension: the spatiotemporal dynamics of word recognition in context. J. Neurosci. 39 (3), 519–527.

Kotz, S.A., Paulmann, S., 2011. Emotion, language, and the brain. Language and Linguistics Compass 5 (3), 108–125. https://doi.org/10.1111/j.1749-818x.2010.00267.x.
Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647.
Leuthold, H., Kunkel, A., Mackenzie, I.G., Filik, R., 2015. Online processing of moral transgressions: ERP evidence for spontaneous evaluation. Soc. Cognit. Affect Neurosci. 10, 1021–1029.
Lu, J., Peng, X., Liao, C., Cui, F., 2019. The stereotype of professional roles influences neural responses to moral transgressions: ERP evidence. Biol. Psychol. 145, 55–61.
Luck, S.J., Hillyard, S.A., 1994. Electrophysiological correlates of feature analysis during visual search. Psychophysiology 31, 291–308.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG-and MEG-data. J. Neurosci. Methods 164 (1), 177–190.
McNamara, T.P., 2005. Semantic Priming. Perspectives from Memory and Word Recognition. Psychology Press Ltd, New York.
Michaels, J.L., Vallacher, R.R., Siebovitch, L.S., 2013. Volatile psychological dynamics in social interactions: attitudes and emotions react asymmetrically to interaction shifts between agreement and disagreement. Social Psychological and Personality Science 4 (6), 705–713. https://doi.org/10.1177/1948550613482985.
Mikhail, J., 2007. Universal moral grammar: theory, evidence and the future. Trends Cognit. Sci. 11 (4), 143–152.
Mikhail, J., 2011. Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment. Cambridge University Press, Cambridge.
Moll, J., de Oliveira-Souza, R., Zahn, R., 2008. The neural basis of moral cognition: sentiments, concepts, and values. Ann. N. Y. Acad. Sci. 1124, 161–180.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J., 2005. Opinion: the neural basis of human moral cognition. Nat. Rev. Neurosci. 6 (10), 799–809.
Neely, J.H., 1991. Semantic priming effects in visual word recognition: a selective review of current findings and theories. In: Besner, D., Humphreys, G.W. (Eds.), Basic Processes in Reading: Visual Word Recognition. Erlbaum, Hillsdale, NJ, pp. 264–336.
Nieuwland, M.S., Van Berkum, J.J.A., 2006. When peanuts fall in love: N400 evidence for the power of discourse. J. Cognit. Neurosci. 18 (7), 1098–1111.
Nohlen, H., van Harreveld, F., Rotteveel, M., Lelieveld, G.-J., Crone, E., 2014. Evaluating ambivalence: social-cognitive and affective brain regions associated with ambivalent decision-making. Soc. Cognit. Affect Neurosci. 9, 924–931. https://doi.org/10.1093/scan/nst074.
Pylkkänen, L., Marantz, A., 2003. Tracking the time course of word recognition with MEG. Trends Cognit. Sci. 7 (5), 187–189.
Rohr, L., Rahman, R.A., 2015. Affective responses to emotional words are boosted in communicative situations. Neuroimage 109, 273–282.
Sarlo, M., Lotto, L., Manfrinati, A., Manfrinati, A., Rumiati, R., Gallicchio, G., Palomba, D., 2012. Temporal dynamics of cognitive–emotional interplay in moral decision-making. J. Cognit. Neurosci. 24 (4), 1018–1029.
Schmidtke, D.S., Schröder, T., Jacobs, A.M., Conrad, M., 2014. ANGST: affective norms for German sentiment terms, derived from the affective norms for English words. Behav. Res. Methods 46 (4), 1108–1118.
Schwarz, N., 2008. Attitude measurement. Attitudes and attitude change. In: Crano, W. D., Prislin, R. (Eds.), 3. Psychology Press, New York, pp. 41–60.
Turner, W.F., Johnston, P., de Boer, K., Morawetz, C., Bode, S., 2017. Multivariate pattern analysis of event-related potentials predicts the subjective relevance of everyday objects. Conscious. Cognit. 55, 46–58. https://doi.org/10.1016/j.concog.2017.07.006.
Van Berkum, J.J., Holleman, B., Nieuwland, N., Otten, M., Murre, J., 2009. Right or wrong? The brain's fast response to morally objectionable statements. Psychol. Sci. 20 (9), 1092–1099.
Võ, M.L.H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M., Jacobs, A.M., 2009. The berlin affective word list reloaded (BAWL-R). Behav. Res. Methods 41 (2), 534–538.
Vogel, E.K., Luck, S.J., 2000. The visual N1 component as an index of a discrimination process. Psychophysiology 37, 190–203.
Wagner, N.-F., Chaves, P., Wolff, A., 2017. Discovering the neural nature of moral cognition? Empirical, theoretical, and practical challenges in bioethical research with electroencephalography (EEG). J. bioeth. Inq. 14 (2), 299–313.
Wing, E.A., Iyengar, V., Hess, T.M., LaBar, K.S., Huettel, S.A., Cabeza, R., 2018. Neural mechanisms underlying subsequent memory for personal beliefs: an fMRI study. Cognit. Affect Behav. Neurosci. 18, 216–231. https://doi.org/10.3758/s13415-018-0563-y.