10.3217/978-3-99161-014-4-054

DECODING MORAL JUDGEMENT FROM TEXT: A PILOT STUDY

Diana E. Gherman[1], Thorsten O. Zander[1]

[1] Brandenburg University of Technology Cottbus–Senftenberg, Germany

E-mail: diana.gherman@b-tu.de

ABSTRACT: Moral judgement is a complex human reaction that engages cognitive and emotional dimensions. While some of the morality neural correlates are known, it is currently unclear if we can detect moral violation at a single-trial level. In a pilot study, here we explore the feasibility of moral judgement decoding from text stimuli with passive brain-computer interfaces. For effective moral judgement elicitation, we use video-audio affective priming prior to text stimuli presentation and attribute the text to moral agents. Our results show that further efforts are necessary to achieve reliable classification between moral congruency vs. incongruency states. We obtain good accuracy results for neutral vs. morally-charged trials. With this research, we try to pave the way towards neuroadaptive human-computer interaction and more human-compatible large language models (LLMs).

INTRODUCTION

Passive BCIs. Passive brain-computer interfaces (pBCIs) can seamlessly decode mental states from a user's brain activity [1]. Active BCIs require the conscious and intentional modulation of one's brain activity, while reactive BCIs make use of external stimuli such as flickering lights to evoke a desired reaction [2]. Meanwhile, pBCIs operate in the background, capturing the spontaneous reactions to specific stimuli in the environment. Most commonly, electroencephalography (EEG) signals are collected and used for mental state classification. Once decoded, pBCIs can provide this real-time information to a computer that can then adapt its outputs to cater to individual needs and preferences. This new form of interaction has previously been described as neuroadaptive [3]. Thus, pBCIs could upgrade human-computer interaction (HCI) to a more natural, fluid type of communication that can be employed in various fields. The potential for safer and more efficient occupational environments through neuroadaptivity has been shown for driving [4], aviation [5] and medicine [6], but also for leisure activities such as gaming [7]. Among others, cognitive states like workload [8], error-perception [9] and surprise [10] have been successfully decoded with pBCI. While extensive research has been done to explore average EEG correlates of emotions, there are relatively few studies that demonstrate robust capabilities for emotional state detection at a single trial level [11,12]. The most common types of features used for emotion classification are event-related potentials (ERPs), frontal EEG asymmetry and event-related desynchronization / synchronization [13]. To investigate single-trial emotion detection from ERPs, a recent study combined workload and stress detection in a social evaluation context [14]. Using a cross-subject classification technique with transfer learning, stress vs. relaxation levels were detected with an average accuracy of over 80%. Single-trial classification of emotion based on ERPs was also achieved for different levels of valence and arousal with a definite advantage for arousal discrimination in [15] and [16]. Another study using EEG recorded while participants were watching music videos managed high classification accuracies for stress levels by using entropy-based features [17]. Our study proposes exploring how well pBCI systems can perform in classifying a specific type of emotion, moral emotion [18]. According to the well-known arousal-valence dimension model of emotions [19], moral violations could evoke high arousal and negative valence emotions such as anger or disgust [20,21]. In contrast, congruent moral stimuli could be associated with low arousal and positive valence. In this investigation, we try to decode moral emotions with pBCI through moral judgements.

Moral judgement. We operationalize here moral judgement as the degree of agreement or disagreement to morally-charged contexts. Moral judgement is a complex human reaction that can include both a cognitive and emotional dimension [22,23]. As an automatic and emotional response, moral judgement can be triggered at an unconscious, intuition-based level, determined by a combination of factors such as personality, culture or motivation [24,25] and is associated with deeper structures of the brain [26]. On the other hand, especially when explicit moral reasoning is required, cognitive functions such as inhibition, cognitive conflict, memory and theory of mind processes are engaged and different prefrontal cortical areas become more active [27,28]. A morally-charged stimulus can either resonate with or challenge an individual's moral perspective, thereby evoking a meaningful moral reaction. This depends on the congruency moral stance with one's personal values and experience with a particular topic. This reaction can be recorded with brain imaging methods such as EEG and potentially decoded with pBCI. While some EEG

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-054

studies looked at the signal patterns associated with neutral, positive, and negative moral judgements, there has not been much work investigating the feasibility of single-trial moral judgement detection for text stimuli [29]. In [30], 90 morally consistent and inconsistent statements were presented to pre-selected groups consisting of Christian and non-Christian male participants while recording their electroencephalography (EEG) data. The statements were displayed one word at a time, with the final word of each determining the overall moral meaning. In reaction to these key words, a small N400 event related-potential (ERP) was found for morally-incongruent words. Also, a late positive potential (LPP) was found around 500-600 ms. The congruency of the moral words was determined based on participants' religiosity for relevant topics (e.g. "I think euthanasia is acceptable/unacceptable"). Another similar study [31] used morally acceptable or unacceptable statements (aligned or misaligned with social norms) presented word by word to elicit moral agreement or disagreement. They also found an LPP around the fronto-parietal region in the case of unacceptable statements. A more recent study that used a multivariate pattern classification (MVPA) showed that agreement or disagreement to morally-charged statements (e.g. "Wars are acceptable / unacceptable") could be predicted from 180ms following the critical ending words, based on the approval or disapproval with these statements indicated via button presses ("yes" and "no") [32]. Moral attitudes regarding particular topics are acquired throughout one's life and are strongly correlated with views and values assimilated within family, society, and personal experiences. The context in which statements appear is also important in eliciting corresponding moral reactions. Previous studies have shown that negative emotion can that trigger a signalling mechanism, making moral situations more salient [22]. Thus, a realistic emotional context used as an affective priming for the textual stimulus could significantly help in this elicitation, as compared to passive statements devoid of context [33,34]. This might be especially relevant for single trial detection. Also, existing theories on effective emotion elicitation attest to the importance of constructing agents for moral assessments to be attributed to, which also improve the elicitation of moral reactions, making the experience more relatable and impactful [35,36]. In this paper, we investigate the feasibility of moral judgement decoding with pBCI for morally-charged statements presented following affective priming represented by emotional videos on specific topics. Previous work has identified video-based stimuli with audios to be considerably more efficient in emotion elicitation, as they are more realistic [37] and produce the highest number of statistically significant features [38]. While most studies that used affective priming in the context of moral judgement assessment so far have used text-based priming, we explore the use of videos with audio here. In light of an increasingly digitized world and advanced artificial intelligence systems (AI) such as large language models (LLMs) [39], successful real-time decoding of moral judgement could open a new realm of possibilities for better and more human-compatible HCI through neuroadaptivity.

MATERIALS AND METHODS

Participants This pilot study included 3 participants (2 males, and 1 female) with a mean age of 31 years. The experimental procedure was approved by the Research Ethics Committee of the Brandenburg University of Technology Cottbus-Senftenberg (ID: EK2024-03).

EEG recording. Their EEG data was recorded using an ActiCHamp amplifier with 64 active actiCAP slim gel electrodes (Brain Products GmbH, Gilching, Germany). The system provides an electrode montage along the extended international 10-20 system (see https://www.brainproducts.com/downloads/cap-montages/ for detailed positions). On the used hardware platform the recorded data is natively reference-free and was common-average referenced after recording. The signal was sampled at 500Hz.

Experiment overview. The task involved watching videos and reading statements related to 4 social justice issues: immigration, racial discrimination, sexism, and homosexuality. Sixteen videos were presented in a random order, followed by 10 randomized statements (5 morally agreeable/congruent and 5 morally disagreeable/incongruent). The utilized videos were collected directly from YouTube or complied together using sequences from a longer Youtube video, such that each video lasted approximately 1 minute. They represented a segment from real TV or media news and they were generally found on channels of multimedia news organizations. Each video included audio as well. After each visualisation, the participants would read an instruction informing them the upcoming statements would be comments left under the respective video by people on the internet. Thus, we are framing strangers on the internet as moral agents responsible for their actions, here agreeable or disagreeable statements. In reality, statements were created by experimenters with the help of the publicly available LLM, ChatGPT. The written statements are presented word by word, in a Rapid Serial Visualization Presentation (RSVP) manner [41] with an Optimal Recognition Point (ORP) alignment of the words [42]. Corresponding with the ORP position, the aligned letters in each word were presented in red, while the rest of the letters remained black. While pre-defined labels were set for these key words, the participants also indicated their agreement, disagreement, or uncertainty regarding the read statement by using keyboard buttons (left arrow for *agree*, downward arrow for *uncertain* and right arrow for *disagree*). The position of the agreement level buttons on the screen corresponded with the position of the response keyboard buttons. The ending, key word of each sentence determined the entire moral stance of the

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-054

sentence and represented the events used for classification. In total, there were 160 statements and therefore, 160 key words in the task: 80 morally-congruent and 80 morally-incongruent. Following, we will illustrate one video-statements example. One of the videos included in the study was a short news piece on the persecution and abuse gay people experience in Uganda due to strict anti-homosexuality laws. In this, there are depictions of people expressing their disappointment and fear regarding these laws and sequences of politicians communicating morally controversial statements such as: "We are going to reinforce the law enforcement officers to make sure that homosexuals have to space in Uganda." After this video the sentences in Tab. 1 were presented in a randomized, word by word manner. The speed of the word presentation differed based on the character length of each non-key word, with 700ms base time and 20ms added for each character besides the first one. For instance, the word *must* was presented for 760ms. The ending, morally-charged words were all presented for 1500ms. The task lasted one hour.

Table 1. Example of morally congruent and incongruent statements

Congruent
1. Uganda's laws for gays should be **humane**.
2. Gay people in Uganda deserve **freedom**.
3. Gay people in Uganda should be treated with **dignity**.
4. Gay Ugandan citizens merit more **respect**.
5. Equality or gay individuals in Uganda is **essential**.
Incongruent
6. In Uganda, laws regarding gays must be **harsh**.
7. Uganda's gay people deserve **prison**.
8. Gay Ugandans should be shown **disrespect**.
9. Ugandan gays merit more **punishment**.
10. For Uganda's gays, equality is **unacceptable**.

Classification method. The classification was performed offline, using MATLAB R2022a (The Mathworks, Inc., Natick, MA, USA) and BCILAB 1.4-devel [43]. Responses that did not align with the predefined classes (*congruent* vs. *incongruent*) were excluded from the classification. Thus, in the sentence "Gay people in Uganda deserve freedom." the pre-defined label for the word freedom was congruent. If the participants pressed on the "disagree" or uncertain buttons instead, this trial was excluded from the classification. We also explored the classification of moral (congruent and incongruent moral combined trials) vs. neutral trials. The neutral trials were categorized based on list of 86 words that appeared within sentences. Examples of neutral words include: "eventually, ultimately, casual, concept, idea, fact". A windowed means approach [44] was used for the feature extraction. The data was bandpass-filtered between 0.1 and 15 Hz. Regularized linear discriminant analysis (LDA) with a (5x5)-fold cross-validation was used for the classification of congruent vs. incongruent trials and moral vs. neutral trials. Epochs of 1 second were extracted with a start time at stimulus onset (key word presentation). We explored two sets of 50 ms time windows in which amplitude is averaged. One set of time windows we used were between 300 and 600 ms after the stimulus, with 6 consecutive time windows. The second set of time windows were set between 400ms and 1000 ms, with 12 consecutive time windows. These windows align with the assumed occurrence of N400 and LPP effects as discussed in [30].

RESULTS

The average classification results on congruent vs. incongruent classes (*CvsI*) and neutral vs. moral (*NvsM*) for both sets of time can be seen in Tab. 2. Only one participant reached classifier significance for the 400-1000 set, with an accuracy of 65%. The chance level in this case is at 57%, which coincides with the associated average accuracy. In contrast, all classifiers for both time window sets reached significance for the neutral vs. moral trials. Averaged ERP potentials for channels Fz and Cz were obtained for both types of classes after independent component analysis (ICA) and non-brain component removal. ERPs for morally congruent vs. incongruent trials are illustrated in *Fig. 1* and ERPs for morally-charged vs. neutral trials are illustrated in *Fig. 2.*

Table 2. Classification results for congruent vs. incongruent (*CvsI*) and neutral vs. moral (*NvsM*) trials

Time windows	TP (%)	TN (%)	Accuracy (%)
300 - 600 CvsI / NvsM	49 / 83	52 / 69	50 / 78
400 - 1000 CvsI / NvsM	59 / 80	54 / 72	57 / 77

TP = True positives (incongruent); TN = True negatives (incongruent)

DISCUSSION

While decoding accuracy for morally congruent and incongruent trials was not successful with this simple approach, we could observe good decoding accuracies for neutral vs. morally-charged words. This was also reflected in the grand-average ERP. Our results are not entirely surprising, given the difficulty of emotion detection from EEG at a single-trial level [12] and the complexity of moral emotions. A recent pBCI investigation [29] also found chance-level results when looking at the potential of single-trial detection for morally acceptable and objectionable trials on data collected in [30] and [31]. However, we found good performance classification for neutral vs. morally-charged trials. We postulate that while the chosen moral words are relevant enough to produce genuine reactions

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-054

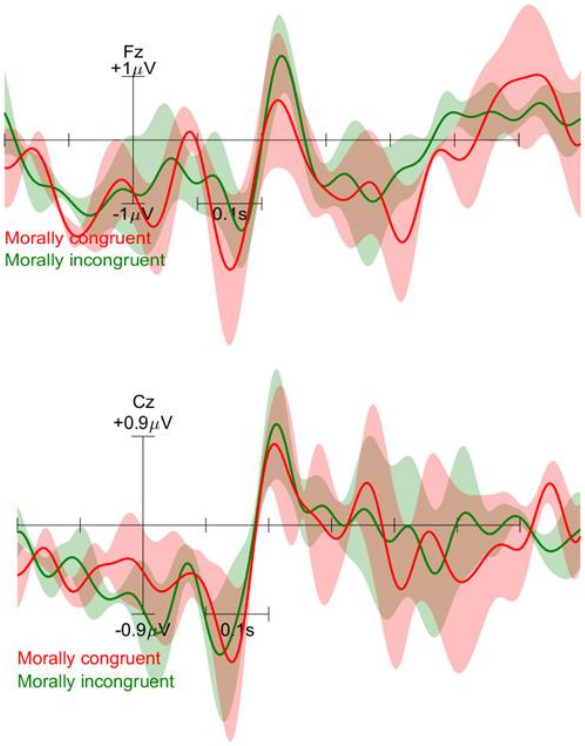

Figure 1. Subject-average ERP potentials for morally congruent and incongruent words.

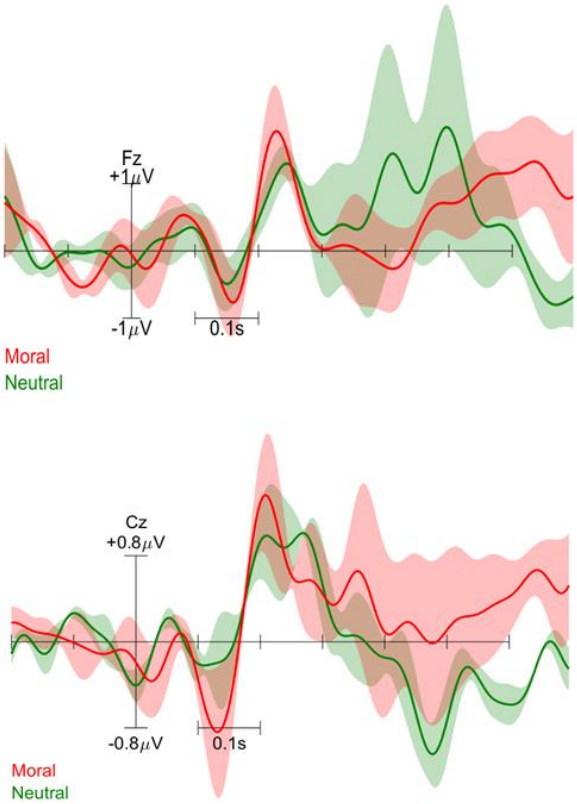

Figure 2 Subject-average ERP potentials for moral and neutral words.

in comparison to neutral stimuli, the current feature extraction and classification approach might need improvements to better capture potential signal differences between morally congruent and incongruent trials. Encouraging results come from recent studies that explored more sophisticated algorithms and feature extraction methods for emotion detection [17,45]. Another way we plan to improve our results in a larger study is to only include participants that align with a certain profile, such that we can ensure they hold clear moral stances towards the topics. Previous studies have identified the importance of moral attitude strength for effective moral emotion elicitation [46] and the corresponding impact on neural signals [23]. In this study we assumed that participants will have the expected, coherent moral value system. We excluded trials, where the manual answers were incompatible with our assumptions. As we only discarded a few trials, we think the participants shown here share our assumed morality. In the recruitment for the main study following up this pilot, we will pre-assess the moral value system of each participant. More specifically, we will include questionnaires meant to assess the participants' attitudes towards sexism [47], immigration [48], racism [49] and homosexuality [50]. Hence, only participants who highly agree with immigration and homosexuality and highly disagree with sexism and racism will be invited to the study. Successful real-time decoding of mental states in reaction to written stimuli could transform human-computer communication in the context of LLMs. For instance, training of LLM could benefit from replacing or augmenting human explicit feedback in Reinforcement Learning with Human Feedback (RLHF) [51,52] with neural-based implicit feedback [53], potentially offering new solutions for a better synergy between humans and machines.

CONCLUSION

In this pilot investigation, we looked at the feasibility of single-trial detection of moral judgement from text after video-based affective priming. Our work offers insights into the neural correlates of moral judgement, as well as ideas for classification improvement for a study that includes more participants and better-suited participant profiles.

REFERENCES

[1] Zander TO, Kothe C. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. J Neural Eng. 2011 Mar 24;8(2):025005.

[2] Wolpaw JR. Chapter 6 - Brain–computer interfaces. In: Barnes MP, Good DC, editors. Handbook of Clinical Neurology. Elsevier; 2013. p. 67–74.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-054

[3] Krol LR, Zander TO. Chapter 2 - Defining neuroadaptive technology: the trouble with implicit human-computer interaction. In: Fairclough SH, Zander TO, editors. Current Research in Neuroadaptive Technology. Academic Press; 2022. p. 17–42.
[4] Lin CT, Wu RC, Liang SF, Chao WH, Chen YJ, Jung TP. EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Trans Circuits Syst I Regul Pap. 2005 Dec;52(12):2726–38.
[5] Dehais F, Dupres A, Di Flumeri G, Verdiere K, Borghini G, Babiloni F, et al. Monitoring Pilot's Cognitive Fatigue with Engagement Features in Simulated and Actual Flight Conditions Using an Hybrid fNIRS-EEG Passive BCI. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE; 2018. p. 544–9.
[6] Zander TO, Shetty K, Lorenz R, Leff DR, Krol LR, Darzi AW, et al. Automated Task Load Detection with Electroencephalography: Towards Passive Brain–Computer Interfacing in Robotic Surgery. J Med Robot Res. 2017 Mar 1;02(01):1750003.
[7] Krol LR, Freytag SC, Zander TO. Meyendtris: a hands-free, multimodal tetris clone using eye tracking and passive BCI for intuitive neuroadaptive gaming. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York, NY, USA: Association for Computing Machinery; 2017. p. 433–7. (ICMI '17).
[8] Gerjets P, Walter C, Rosenstiel W, Bogdan M, Zander TO. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. Front Neurosci [Internet]. 2014;8. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2014.00385
[9] Parra LC, Spence CD, Gerson AD, Sajda P. Response error correction--a demonstration of improved human-machine performance using real-time EEG monitoring. IEEE Trans Neural Syst Rehabil Eng. 2003 Jun;11(2):173–7.
[10] Pawlitzki, J., Klaproth, O., Krol, L.R. and Zander, T.O., 2021. Automation surprise in the neuroadaptive cockpit. In Neuroergonomics Conference.
[11] Alarcao SM, Fonseca MJ. Emotions recognition using EEG signals: A survey. IEEE Trans Affect Comput. 2019 Jul 1;10(3):374–93.
[12] Alimardani M, Hiraki K. Passive brain-computer interfaces for enhanced human-robot interaction. Front Robot AI. 2020 Oct 2;7:125.
[13] Al-Nafjan A, Hosny M, Al-Ohali Y, Al-Wabil A. Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review. NATO Adv Sci Inst Ser E Appl Sci. 2017 Dec 1;7(12):1239.
[14] Bagheri M, Power SD. Simultaneous Classification of Both Mental Workload and Stress Level Suitable for an Online Passive Brain-Computer Interface. Sensors [Internet]. 2022 Jan 11;22(2). Available from: http://dx.doi.org/10.3390/s22020535
[15] Mathieu NG, Bonnet S, Harquel S, Gentaz E, Campagne A. Single-trial ERP classification of emotional processing. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE; 2013. p. 101–4.
[16] Liu YH, Wu CT, Kao YH, Chen YT. Single-trial EEG-based emotion recognition using kernel Eigen-emotion pattern and adaptive support vector machine. Annu Int Conf IEEE Eng Med Biol Soc. 2013;2013:4306–9.
[17] Gao Y, Wang X, Potter T, Zhang J, Zhang Y. Single-trial EEG emotion recognition using Granger Causality/Transfer Entropy analysis. J Neurosci Methods. 2020 Dec 1;346(108904):108904.
[18] Haidt J. The Moral Emotions. In: Handbook of Affective Sciences. Oxford University PressNew York, NY; 2002. p. 852–70.
[19] Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161.
[20] Hutcherson CA, Gross JJ. The moral emotions: a social-functionalist account of anger, disgust, and contempt. J Pers Soc Psychol. 2011 Apr;100(4):719–37.
[21] Chapman HA, Kim DA, Susskind JM, Anderson AK. In bad taste: evidence for the oral origins of moral disgust. Science. 2009 Feb 27;323(5918):1222–6.
[22] Decety J, Michalska KJ, Kinzler KD. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cereb Cortex. 2012 Jan;22(1):209–20.
[23] Hundrieser M, Stahl J. How attitude strength and information influence moral decision making: Evidence from event-related potentials: ERPs in moral decision. Psychophysiology. 2016 May 1;53(5):678–88.
[24] Haidt J. The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev. 2001 Oct;108(4):814–34.
[25] Gaertner SL, McLaughlin JP. Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics. Soc Psychol Q. 1983;46(1):23–30.
[26] Cunningham WA, Raye CL, Johnson MK. Implicit and explicit evaluation: FMRI correlates of valence, emotional intensity, and control in the processing of attitudes. J Cogn Neurosci. 2004 Dec;16(10):1717–29.
[27] Fede SJ, Kiehl KA. Meta-analysis of the moral brain: patterns of neural engagement assessed

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-054

using multilevel kernel density analysis. Brain Imaging Behav. 2020 Apr;14(2):534–47.
[28] Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. The neural bases of cognitive conflict and control in moral judgement. Neuron. 2004 Oct 14;44(2):389–400.
[29] Andreeßen LM. Towards real-world applicability of neuroadaptive technologies: investigating subject-independence, task-independence and versatility of passive brain-computer interfaces [Internet]. BTU Cottbus-Senftenberg; 2023. Available from: https://opus4.kobv.de/opus4-btu/files/6652/Andreessen_Lena.pdf
[30] Van Berkum JJA, Holleman B, Nieuwland M, Otten M, Murre J. Right or wrong? The brain's fast response to morally objectionable statements: The brain's fast response to morally objectionable statements. Psychol Sci. 2009 Sep;20(9):1092–9.
[31] Leuthold H, Kunkel A, Mackenzie IG, Filik R. Online processing of moral transgressions: ERP evidence for spontaneous evaluation. Soc Cogn Affect Neurosci. 2015 Aug;10(8):1021–9.
[32] Hundrieser M, Mattes A, Stahl J. Predicting participants' attitudes from patterns of event-related potentials during the reading of morally relevant statements - An MVPA investigation. Neuropsychologia. 2021 Mar 12;153:107768.
[33] Demel R, Waldmann M, Schacht A. The Role of Emotions in Moral Judgements: Time-resolved evidence from event-related brain potentials [Internet]. bioRxiv. bioRxiv; 2019. Available from: https://www.biorxiv.org/content/10.1101/541342v1.full.pdf
[34] Greenaway KH, Kalokerinos EK, Williams LA. Context is Everything (in Emotion Research). Soc Personal Psychol Compass. 2018 Jun;12(6):e12393.
[35] Gray K, Wegner DM. Dimensions of moral emotions. Emot Rev. 2011 Jul 28;3(3):258–60.
[36] Pantazi M, Struiksma M, Van Berkum J. An EEG study on the role of perspective-taking in the assessment of value-loaded statements [Internet]. Available from: https://studenttheses.uu.nl/bitstream/handle/20.500.12932/11511/Thesis_Pantazi.pdf?sequence=1&isAllowed=y
[37] Rahman MM, Sarkar AK, Hossain MA, Hossain MS, Islam MR, Hossain MB, et al. Recognition of human emotions using EEG signals: A review. Comput Biol Med. 2021 Sep;136(104696):104696.
[38] Masood N, Farooq H. Comparing neural correlates of human emotions across multiple stimulus presentation paradigms. Brain Sci. 2021 May 25;11(6):696.
[39] Levy S. What OpenAI Really Wants. Wired [Internet]. 2023 Sep 5 [cited 2023 Sep 13]; Available from: https://www.wired.com/story/what-openai-really-wants/
[40] Kothe C, Medine D, Grivich M. Lab streaming layer (2014). URL: https://github com/sccn/labstreaminglayer.
[41] Potter M. Rapid serial visual presentation (rsvp): a method for studying language processing. 2018 Apr 17; Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9780429505379-5/rapid-serial-visual-presentation-rsvp-mary-potter
[42] Brysbaert M, Nazir T. Visual constraints in written word recognition: evidence from the optimal viewing-position effect. J Res Read. 2005 Aug;28(3):216–28.
[43] Kothe CA, Makeig S. BCILAB: a platform for brain–computer interface development. J Neural Eng. 2013 Aug 28;10(5):056014.
[44] Blankertz B, Lemm S, Treder M, Haufe S, Müller KR. Single-trial analysis and classification of ERP components--a tutorial. Neuroimage. 2011 May 15;56(2):814–25.
[45] Arjun, Rajpoot AS, Panicker MR. Subject independent emotion recognition using EEG signals employing attention driven neural networks. Biomed Signal Process Control. 2022 May 1;75:103547.
[46] Ugazio G, Lamm C, Singer T. The role of emotions for moral judgements depends on the type of emotion and moral scenario. Emotion. 2012 Jun;12(3):579–90.
[47] Glick P, Fiske ST. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. J Pers Soc Psychol. 1996;70(3):491–512.
[48] Pratto F, Sidanius J, Stallworth LM, Malle BF. Social dominance orientation: A personality variable predicting social and political attitudes. J Pers Soc Psychol. 1994 Oct;67(4):741–63.
[49] Mcconahay JB. Modern racism, ambivalence, and the Modern Racism Scale. 1986; Available from: https://psycnet.apa.org/record/1986-98698-004
[50] Fisher TD, Davis CM, Yarber WL. Handbook of Sexuality-Related Measures. Routledge; 2013. 680 p.
[51] Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, et al. Learning to summarize from human feedback [Internet]. arXiv [cs.CL]. 2020. Available from: https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf
[52] Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, et al. Open problems and fundamental limitations of reinforcement learning from human feedback [Internet]. arXiv [cs.AI]. 2023. Available from: http://arxiv.org/abs/2307.15217
[53] Xu D, Agarwal M, Fekri F, Sivakumar R. Playing Games with Implicit Human Feedback.

Published by
Verlag der Technischen Universität Graz