The time-course of moral perception: An ERP investigation of the moral pop-out effect

Ana Gantman[1], Sayeed Devraj-Kizuk[2], Peter Mende-Siedlecki[3], Jay J. Van Bavel[4], Kyle E. Mathewson[2,5]

[1]Department of Psychology, Brooklyn College (CUNY)
[2]Neuroscience and Mental Health Institute, University of Alberta
[3] Department of Psychological and Brain Sciences, University of Delaware
[4]Department of Psychology, New York University
[5]Department of Psychology, University of Alberta

Corresponding Author: Ana Gantman, Department of Psychology, 2900 Bedford Avenue, Brooklyn, NY, ana.gantman@brooklyn.cuny.edu

Acknowledgements: This work was supported by a Natural Science and Engineering Research Council of Canada discovery grant (#04792) to Kyle E. Mathewson and a National Science Foundation Grant to Jay J. Van Bavel **(#1349089**). The authors would like to thank Tyler Harrison and Brian Steele for assistance with data collection and members of the Social Perception and Evaluation lab for comments on this manuscript. This research was presented at Cognitive Neuroscience Society by PMS, Brian Steele at the 34th Banff Annual Seminar in Cognitive Sciences and the Royce Psychology Research Conference. Responsibilities: AG, PMS, JVB, KEM designed the experiments, SDK and AG analyzed the data with input from PMS, JVB, and KEM. AG, PMS, JVB, and KEM wrote the manuscript.

Abstract

Humans are highly attuned to perceptual cues about their values. A growing body of evidence suggests that people selectively attend to moral stimuli. However, it is unknown whether morality is prioritized early in perception or much later in cognitive processing. We use a combination of behavioral methods and electroencephalography to investigate how early in perception moral words are prioritized relative to non-moral words. The behavioral data replicate previous research indicating that people are more likely to correctly identify moral than non-moral words in a modified lexical decision task. The electroencephalography data reveal that words are distinguished from non-words as early as 200 milliseconds after onset over frontal brain areas, and moral words are distinguished from non-moral words 100 milliseconds later over left-posterior cortex. Further analyses reveal that differences in brain activity to moral vs. non-moral words cannot be explained by differences in arousal associated with the words. These results suggest that moral content might be prioritized in conscious awareness after an initial perceptual encoding but before subsequent memory processing or action preparation. This work offers a more precise theoretical framework for understanding how morality impacts vision and behavior.

Keywords: morality, EEG, social neuroscience, conscious awareness, vision

The time-course of moral perception: An ERP investigation of the moral pop-out effect

Morality is such an integral part of social life that we must be vigilant for cues about the moral values in our group. Yet, most models of moral psychology are based on tasks in which people read (and must understand) moral dilemmas (e.g., Greene & Haidt, 2002; Haidt, 2001; Kohlberg, 1979; Paxton & Greene, 2010; for a recent review Everett & Kahane, 2020). This approach has proven fruitful, but it overlooks an important pre-cursor to moral judgment and decision-making: how do people *see* a moral stimulus in the first place? To better understand the precursors to moral judgment, a small but growing body of research has begun to examine how morality shapes vision and, in turn, is shaped by vision (Gantman & Van Bavel, 2014, 2015a, 2016). The current research uses a combination of behavioral and neuroscientific methods to investigate exactly when in the perceptual processing stream moral information is treated differently than other information.

Morality plays a central role in the life of social groups (Haidt, 2008). To signal their alignment with their group (Haidt & Graham, 2009) and avoid violating group values that may bar them from access to key social and psychological resources, individuals need to remain vigilant for morally relevant information in the environment. Indeed, evidence suggests that people have heightened awareness of moral stimuli (Gantman & Van Bavel, 2015a)—especially when moral goals are active (Gantman & Van Bavel, 2016). Moreover, people show enhanced perception for the faces of people who have committed morally bad actions (Anderson, Siegel, Bliss-Moreau, & Barrett, 2011) as well as visual depictions of bad outcomes befalling morally bad actors (Callan et al., 2013). Because the ability to recognize moral situations and act appropriately is critical to one's status in social groups, people may be highly attuned to the presence of moral stimuli.

Taken together, this recent work (e.g., Gantman & Van Bavel, 2015a) suggests that morality may shape numerous aspects of perception—from visual attention to conscious awareness. However, the link between morality and visual experience is still underspecified. The studies to date largely focus on explicit behavioral responses to moral stimuli in order to make inferences about the underlying processes. For example, in one set of tasks, participants were presented with backward masked moral, non-moral words, and scrambled non-words in a lexical decision task (Gantman & Van Bavel, 2014). When asked to judge whether a given stimulus (presented at the threshold for conscious awareness) was a word or not, participants were correctly identified moral words with a higher degree of accuracy than non-moral words. This phenomenon—whereby perceivers demonstrate a heightened awareness of morally-laden content—has been termed the moral pop-out effect.

The time-course of moral perception remains uncertain, however. In the case of the moral pop-out effect, for example, it is unclear how early in processing moral words are differentiated from non-moral words (e.g., in perception vs. memory; Firestone & Scholl, 2014; Gantman & Van Bavel, 2015; Firestone & Scholl, 2016; Gantman & Van Bavel, 2016). For instance, some work has argued that visual perception is impenetrable to the influence of morality (Firestone & Scholl, 2014)—part of a longstanding debate in cognitive science about cognitive penetrability (Firestone & Scholl, 2016). Therefore, more work is needed to understand exactly when in perceptual processing moral content is prioritized. We aimed to clarify this process by using a measure with precise temporal resolution. This will allow us to develop a more refined theoretical understanding of the relationship between vision and morality.

As such, the current research examines how early morality is prioritized. We used electroencephalography (EEG) because this method offers highly precise temporal resolution—

at the order of milliseconds—to interrogate the timing of moral information processing (Luck, 2005; Luck, & Kappenman, 2011). For instance, several studies using event-related potentials (ERPs) have found that social categories can influence perceptual processing very quickly (Ito & Cacioppo, 2000; Smith et al., 2003; Ito et al., 2007). In this way, ERPs can be thought of as an additional reaction time measure of the mental computations that unfold between the stimulus and response. As such, temporal distinctions between deflections in the ERP signal are meaningful, and may reflect different stages of perceptual and cognitive processing. The amplitude of these deflections can be used to estimate the amount of or ease of processing at each stage of processing. In this way, the study of ERPs provides a lens into the implicit perceptual and cognitive processing that is occurring online while people are completing a task (Cunningham, Packer, Kesek, & Van Bavel, 2009).

Indeed, recent work has used ERP analyses to examine the temporal dynamics underlying moral judgment and evaluation. In such studies, participants were recorded while processing morality-related words (e.g., Yang, Luo, & Zhang, 2017), behaviors (e.g., Yang, Li, Xiao, Zhang, & Tian, 2014), and images (e.g., Decety & Cacioppo, 2012), often in direct comparison to core disgust-related stimuli (e.g., Yang et al., 2014, 2017). Taken together, the extant ERP literature on morality is somewhat mixed: across this work, various researchers have observed morality-related changes in amplitudes in the N1 (Gui, Gan, & Liu, 2016; Yoder & Decety, 2014), N180 (successful versus attempted harm/help; Gan et al., 2016), recognition potential (versus neutral words; Yang et al., 2017), P200 (shame versus guilt; Zhu et al., 2019); P300 (versus neutral behaviors; Yang et al., 2014), N400 (versus neutral words; Luo et al., 2013), and late positive potential (LPP; Gui et al., 2016; Leuthold, Kunkle, Mackenzie, & Filik, 2015). Moreover, this work indicates that various ERP components related to morality are sensitive to

both valence and arousal (e.g., valence: N1 & N2, Yoder & Decety, 2014; LPP, Leuthold et al., 2015; arousal: N2, Gui et al., 2016; P200, Sarlo et al., 2012; LPP, Gui et al., 2016; Yoder & Decety, 2014). The current paper builds on this prior work by comparing moral to non-moral information processing to better understand the time course of brain activity between stimulus and response that leads to a response advantage for moral words. Given the relative heterogeneity of this previous work, the present work will serve to add clarity to the existing ERP literature on morality.

We reasoned that using an implicit measure of information processing is critical to this particular issue since behavioral responses can be easily contaminated with attention, memory, and motor responses. This makes it very difficult to draw any firm conclusions about the underlying perceptual processing facilitating the moral pop-out effect. Using EEG allows us to determine when moral and non-moral stimuli are differentiated in the brain—even if this process is happening outside conscious awareness or before people can register a behavioral response to a stimulus. This approach also reduces any concerns about demand effects since people are unable to generate specific patterns of brain activity in the same way they might be able to perform a specific pattern of behavior. We see three plausible alternatives for when moral content influences brain activity. We present them here in chronological order for clarity:

Hypothesis 1: Morality affects early visual representations

There are multiple possible times at which moral information might begin to alter stimulus processing. One extreme possibility is that the presence of morally relevant stimuli could affect very early visual representations of the stimuli—possibly in the first few hundred milliseconds of visual processing, when the bottom-up information is transmitted up the ventral visual pathway to identify the stimulus. Another possible pathway through which moral

relevance could affect early visual presentations would be if morally relevant stimuli bypass visual cortex and transmit directly to subcortical structures like the amygdala (Garrido, Barnes, Sahani, & Dolan, 2012; Tamietto & de Gelder, 2010). For instance, there is some evidence that early sensory ERP components like the P1, N1, P2, and N2 can be influenced by both top-down factors, like the direction of spatial attention (Zhang and Luck, 2008) and cognitive control (for a review, Folstein & Van Petten, 2008), as well as bottom-up factors, like stimulus salience (e.g. Strayer and Johnston, 2000). Indeed, the rapid aspects of social perception can be shaped by motivational states (Cunningham et al., 2014), including morally relevant information (Gui et al., 2016; Yoder & Decety, 2014). Thus, it seems possible that differences in the prioritization of moral words may occur during this very early perceptual stage, which can be influenced by higher order states.

It might also be possible to find evidence for morality affecting early visual representations if we saw a difference in brain activity for moral (vs non-moral) stimuli emerge prior to differences between words and non-words. This is theoretically possible, though slightly difficult to make sense of in the domain of words; the moral relevance of the word would have to come through before evidence of recognizing the letter string as a word. This may be more plausible in a different domain (e.g., ambiguous objects) where the moral relevance might be ascertainable prior to precise object recognition (e.g., a burnt piece of toast could be evaluated as morally relevant before people consciously discern an image of Mother Teresa or Jesus). Instead, to find evidence of moral relevance affecting brain activity in the earliest response window, we reasoned that early processing of words (vs. non-words) could happen concurrently with moral (vs. non-moral) processing. Word vs. non-word processing must unfold over the next few milliseconds after stimulus onset, and we could see differences in brain activity to moral (vs.

non-moral) stimuli occurring in this same window (i.e., at the same time as the word vs. non-word processing).

Hypothesis 2: Morality affects what reaches perceptual awareness

A second possibility is that moral stimuli are prioritized after initial perceptual encoding but before memory or response preparation, enhancing perceptual awareness of these stimuli. The P3 is thought to represent a *post-perceptual* cognitive process associated with the transition from perception into response preparation elicited by task relevant and motivationally significant stimuli that emerges as early as 250 after stimulus onset (Squires, Donchin, Herning, & McCarthy, 1977; Polich, 2012). For instance, the P3 is larger for motivationally significant stimuli (e.g. Johnston, Miller, and Burleson, 1986), including one's own name (i.e., the "cocktail party effect"; Gray et al., 2004). According to one popular account, the P3 has been proposed to represent a physiological ignition event associated with widespread neurotransmitter release and communication of the perceptual information to support response preparation and conscious awareness (Dehaene et al., 2006). If so, this would occur after the initial few hundred milliseconds of visual processing—shortly after the brain distinguishes the nature of the stimuli (e.g., word vs. non-word), but before subsequent cognitive and motor processing represented by the LPP. However, enhanced awareness of moral stimuli, would likely be followed by downstream effects during response preparation. In other words, morally relevant stimuli could produce a kind of cocktail party effect; instead of suddenly hearing one's own name "pop out" among the party noise, one might hear a morally relevant word or phrase.

Hypothesis 3: Morality only affects cognitive processing

A third possibility is that moral content influences brain activity during much later, cognitive processing. According to this account, differences between moral and non-moral

stimuli may reflect differences in cognition (e.g., memory retrieval), or in preparing a response to the stimulus (e.g. motor facilitation). For instance, the late positive waveform starts 400 or more milliseconds after stimulus presentation and is a long-lasting wave form that often persists for 500 milliseconds (Friedman & Johnson, 2000). This waveform reflects substantial cognitive processing associated with response preparation, which is sustained long after the initial presentation of a stimulus. When individuals dedicate increased mental resources to processing a stimulus and preparing a response, an increase in this late positive waveform is observed (Gliden, Vaughan, & Costa, 1996; Schupp et al., 2000), including in the context of moral judgment (Leuthold et al., 2015; Gui et al., 2016). If we do not observe a difference between moral and non-moral stimuli until this later time window, it would provide evidence suggesting that the response advantage for morally relevant stimuli reflects additional cognitive processing following initial perception.

Overview of Current Research

In the current research, we measured ERPs evoked by moral and non-moral words to measure how early in the visual processing stream moral words bias brain activity. *When* the brain distinguishes between moral and non-moral classes of stimuli will provide adjudicating evidence between these three different explanations: early in perception, intermediary at the gateway to consciousness in the P3 time window, or late in response preparation in the late positive window.

Previous research demonstrates that people are more likely to correctly identify rapidly presented moral words compared to matched non-moral words (Gantman & Van Bavel, 2014; 2015), but only when the words were presented near the threshold for perceptual awareness—a phenomenon known as the moral pop-out effect. Using the same approach, we presented moral

and non-moral words very rapidly in a lexical decision task. The current research sought to determine when in the visual processing stream moral words are differentiated from non-moral words. This gave us the opportunity to pursue three research goals at once. First, we replicated the moral pop-out effect behaviorally, and examined precisely when moral words were prioritized in the perceptual processing stream; Second, we used this method to examine if moral relevance affects visual experience at the level of perception, awareness, or only later in memory (for debate see Gantman & Van Bavel, 2015a; Firestone & Scholl 2015b; Gantman & Van Bavel, 2015c). Third, we have contributed to the growing literature investigating morality related changes in brain activity (as measured by ERPs; Luck, 2005; Luck, & Kappenman, 2011; Yang, Luo, & Zhang, 2017; Yang, et al, 2014; Decety & Cacioppo, 2012; Gui, Gan, & Liu, 2016; Yoder & Decety, 2014; Gan et al., 2016; Zhu et al., 2019; Luo et al., 2013; Leuthold, Kunkle, Mackenzie, & Filik, 2015).

If the moral pop-out effect is entirely dependent on low-level visual perception via very early processing, then it would be predicted that moral words would differ in the first 100 ms of visual processing, as the bottom-up information is transmitted up the ventral visual pathway to identify the word (Luck & Kappenman, 2011). If we do not see differences until well after 400 ms, then the moral pop-out effect would be better understood as a cognitive or response preparation effect, rather than a perceptual phenomenon *per se* (Luck & Kappenman, 2011). However, if the differences in processing emerge between these very early and late waveforms, this would suggest that moral words may receive a preferential gate into conscious awareness, perhaps due to their motivational relevance (Koudier et al., 2013; Gantman & Van Bavel, 2016).

Methods

Participants

A total of 54 paid volunteers (mean age = 23.26; 36 female, 1 non-binary) from the University of Alberta community gave informed consent, as approved by the internal Research Ethics Board. We used the same sample size from previous work on the moral pop-out effect, as the task participants completed was almost identical (Gantman & Van Bavel, 2014). The experimental design was fully within-subjects to optimize statistical power. All participants reported normal or corrected vision and English as their first language. There were no other exclusionary factors. They were compensated with a $10 honorarium for their participation. Of the initial 54 participants, four individual results were removed from the data before analysis because one did not respond on the majority (74%) of trials, one had a corrupted data file, and two highly favored the non-word response (100% and 93%). (As the overall distribution of trials was 50% word, 50% non-word, and presentation was manipulated to achieve perceptual ambiguity [75% overall accuracy; see below], this non-differentiation might indicate that these latter participants simply did not adequately see the presented stimuli, or simply were not compliant with the task instructions.) This left a final sample of 50 participants for our within-subjects experiment. We report how we determined our sample size, all data exclusions, all manipulations and all measures in the study.

Materials and Procedure

A slightly modified version of the lexical decision task from Gantman and Van Bavel (2014) was employed. Participants sat in a dark room, 57 cm from a 1920 x 1090 pixel^2 (22.5" diagonal) ViewPixx/EEG LCD monitor (VPixx Technologies) with a refresh rate of 120 Hz. Stimuli were presented on a 50% grey background with a Windows 7 PC running Matlab R2012b with the Psychophysics toolbox (Version 3.0.14; Brainard, 1997). Video output was sent to the ViewPixx/EEG with an Asus Striker GTX760 graphics processing unit. The experimental

task script, all analysis scripts, and all data files are available upon publication or during review at the Open Science Foundation (https://osf.io/5jmze/) and Github (https://github.com/kylemath/MoralWordEEG).

On each trial, participants were instructed to fixate on a central cross, attend to the lexical stimulus being presented, and determine whether they had been presented with a word or a non-word (Figure 1). The black central fixation cross was present for a variable pre-trial fixation period between 400 and 700 ms, then the target letter string was presented at fixation for 16.66 ms. The letter strings were presented in 8% gray in font size 20 against a 50% gray background. The target was followed by the fixation cross for a fixed 33.33 ms period, giving a stimulus onset asynchrony (SOA) of 50 ms between the target and the backward mask (a string of ampersands as long as the preceding target word), which was presented for 25 ms. The masks were presented in 30% gray in font size 20 against the same 50% gray background. Font size, background color, and stimulus duration were determined through piloting to ensure that stimuli would be perceptually ambiguous, such that overall accuracy was near 75%, 50% would represent completely random responding, and 100% would represent complete accuracy). Previous work on the moral pop-out effect has shown that differences in moral and non-moral word detection occur only when words are presented ambiguously (Gantman & Van Bavel, 2014; 2015). As such, we decided *a priori* to present stimuli that would be perceptually ambiguous (i.e., correctly identified 75% of the time), which we accomplished by manipulating font size, background color, and stimulus duration.

Participants were instructed to attempt to discriminate whether a word or a non-word was presented. Following mask offset, participants responded to the target by pressing the "1" key (to indicate seeing a word) with their left index finger or by pressing the "5" key (to indicate seeing

a non-word) with their right index finger. In the event that they were unsure of the target, participants were instructed to guess. The next pre-trial fixation period began after a fixed response period of 1500 ms after mask offset, regardless of whether the participant responded.[1]

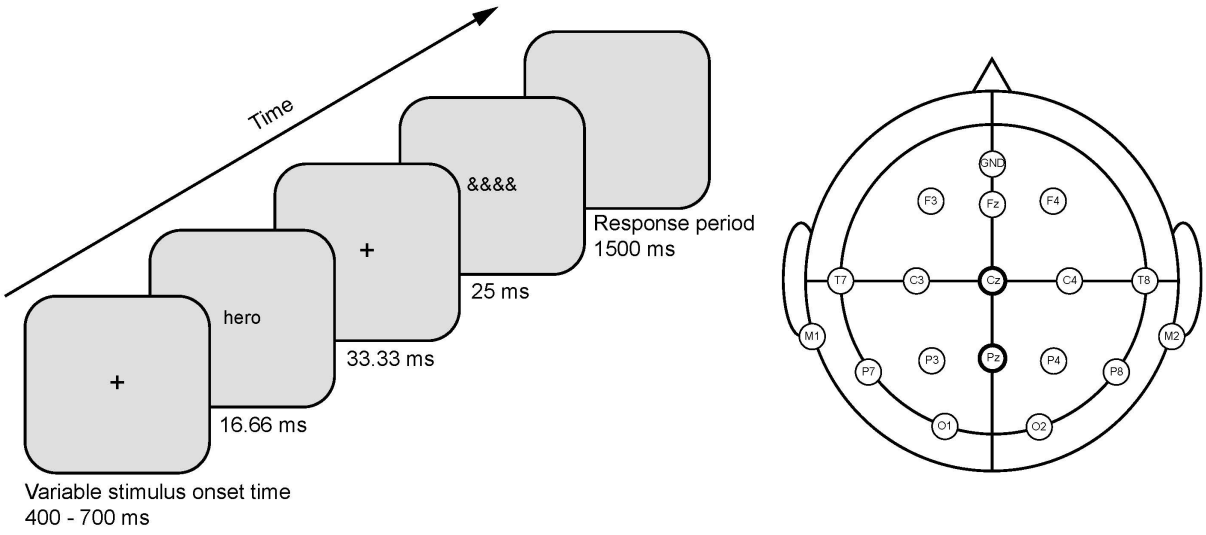

Figure 1. *Schematic of lexical decision task. Participants saw a fixation cross for a variable period, followed by either a moral word, non-moral word, or scrambled non-word displayed for 16.6 ms. After a 50ms SOA, a backward mask was presented for 25ms. The screen remained blank for 1500 ms during which participants indicated with a '1' or '5' whether the string of letters appeared as a word or a non-word, respectively. Stimuli are not shown to scale. Scalp topography shows the layout of the 15 scalp electrodes, the ground at AFz, and 2 references on the mastoids. Darker circles indicate the electrodes from which data are presented.*

Participants completed 10 blocks of 50 trials. Trial order was pseudo-randomized. Our stimuli comprised 125 moral words (e.g., *kill, should*), 125 non-moral words (e.g., *die, could*), and 250 non-words. Words were taken from the initial moral-word pop-out study (Gantman & Van Bavel, 2014) and a slightly modified replication (Firestone & Scholl, 2014). Non-words were scrambled versions of the moral and non-moral words to keep overall letter content between words and non-words constant. Words were length- and frequency-matched with the

[1] In previous research, the next trial would not advance until a response was made. When trials are removed in which participants made no response the moral pop-out effect is unchanged, $B = .17$, $SE = .05$, 95% CI [.08, .26], $p < .005$, $z = 3.60$.

Corpus of Contemporary American English (COCA; Davies, 2008). A meta-analysis of previous studies found that the moral pop-out effect remained significant even after adjusting for any differences in valence, extremity, and reported arousal between moral and non-moral words (Gantman & Van Bavel, 2014). Further analyses regarding valence and arousal are also included in this paper. We note that while the present sample of participants did not provide their own ratings of the moral relevance of the words in our stimulus set, subjective ratings acquired across multiple sets of previous participants established the construct validity of this manipulation (Gantman & Van Bavel, 2014; Firestone & Scholl, 2014).

Prior to the experiment participants performed a practice block with 50 words and 50 non-words, all non-moral and different from those used in the main study. Practice words were presented at high contrast (black) in order to familiarize the participants with the word types (e.g., words vs. non-words) with full awareness of them, but all other materials and timing were identical.

Following the experiment, participants were sent an electronic questionnaire that assessed their belief in a just world (Lerner & Miller, 1978), moral identity (Aquino & Reed, 2002), political ideology, and trust in the justice system (devised by the authors). However, as not all participants completed this survey, these data will not be further analyzed in the present manuscript, but they are available on OSF for secondary analyses (https://osf.io/5jmze/).

Analytic Strategy for Lexical Decision-Task

We used generalized estimating equations (GEE) to estimate our regression parameters instead of ordinary least-squares regression (Zeger & Liang, 1986). This allowed us to take learning effects and other forms of interdependence among participants' responses into account without assuming homogeneity of variance (see also Gantman & Van Bavel, 2014; 2015).

Because our stimuli were presented in random order, an exchangeable correlation matrix was specified for all models (Ballinger, 2004). For analyses using GEE models, we report unstandardized regression coefficients (*B*), standard errors (*SE*) and Wald *Z*'s (for a similar analytic strategy, see Stern, West, Jost, & Rule, 2013; Freeman, Johnson, Ambady, & Rule, 2010). Moral words were coded as 1, non-moral words were coded as 0. To provide further information about effect size, 95% confidence intervals on *B* values are also reported. When analyzing the behavioral results, we included all trials, as the EEG artifacts do not interfere with our behavioral measure. (However, when analyzing ERP results, trials with artifacts were trimmed from the data; see *ERP Preprocessing*.)

EEG Recording

EEG was recorded from 15 scalp locations (O1, O2, P7, P3, Pz, P4, P8, T7, C3, Cz, C4, T8, F3, Fz, F4; 10/20 system) and the right mastoid, referenced online to an electrode affixed to the left mastoid, with a ground channel at AFz, using 18 Ag/AgCl sintered ring electrodes (EasyCap) in a 20-channel electrode cap (EasyCap; Figure 1B). The voltage differences were amplified with a 16-channel V-amp amplifier (Brain Products). Impedance between electrode and scalp was reduced using both abrasive tape and Abralyt gel, until each electrode had impedance below 10 kOhms. Electrode locations were re-referenced offline to the average of the left and right mastoids. The bipolar vertical and horizontal electrooculogram (EOG) was recorded with additional electrodes using two BIP2AUX converters in the V-amp auxiliary channels (Brain Products). Electrodes were placed 1-cm lateral from the outer canthus of each eye, and above and below the left eye. These EOG electrodes had their own ground affixed in the central forehead.

Data were recorded at 1000 Hz with a resolution of 24 bits (0.049 uV steps), and were filtered online with a 0.1 Hz high-pass filter and 200 Hz low-pass filter. Data was collected inside a sound and radio-frequency attenuated chamber (40A-series; Electro-Medical Instruments), with copper mesh covering a window. The lights were off during the experiment, and the chamber window was covered. The only electrical devices inside the chamber were the amplifier (powered from a battery powered laptop located outside the chamber), speakers, keyboard, and mouse (all three powered from outside of the room), the ViewPixx monitor, powered with DC power from outside the chamber, and a battery-powered intercom. There was nothing connected to the internal power outlets. All electrical devices (e.g., cellphones) were removed from the chamber before recording.

EEG Preprocessing

All analyses were completed in Matlab R2012b using the EEGLAB toolbox (Version 13.3.2b; Delorme & Makeig, 2004), as well as custom scripts (https://github.com/kylemath/MoralWordEEG). For ERP analysis, data were filtered offline with a 30 Hz low-pass FIR filter (eegfilt() in EEGlab). Trials were epoched into 2000 ms segments locked to the onset the stimulus, including 1000 ms before the cue as a baseline period. The start of each trial was centered around 0μV by subtracting the average voltage in the 200 ms of the baseline immediately prior to the cue, on each trial and for every electrode. To remove large artifacts due to movement or other non-cognitive factors, trials with absolute voltage fluctuations on any channel greater than 1000 μV were discarded. Eye movements were then removed from the data using the regression-based eye-movement correction procedure developed by Gratton, Coles, and Donchin (1983). After identifying blinks with a template-based approach, this technique computes propagation factors as regression coefficients predicting the vertical and

horizontal eye channel data from the signals at each electrode. The eye channel data is then subtracted from each channel, weighted by these propagation factors, thus removing any variance in the EEG predicted by eye movements recorded in the EOG. Finally, after a second baseline subtraction with the 200 ms pre-cue (since the eye-correction procedure can shift the baseline as well), trials with remaining absolute voltage fluctuations on any channel greater than 500 μV were removed from further analysis. An average of 420.28 (*SD* = 93.01) of the original 500 trials (i.e, 84%) remained for analysis after artifact rejection.

ERP Analysis

The average voltage across all 2000 ms epochs that survived artifact rejection was computed within each condition, for every electrode, and every participant. We restricted our ERP analyses to trials where the target was correctly identified as a word or a non-word, since we cannot be sure what participants saw on incorrect trials. ERPs for each individual were computed by averaging across correct trials within each electrode and each condition. These ERPs were then plotted as a grand average across all electrodes and all participants. In addition to the grand-average ERPs within each condition, we also computed the within-subjects difference for four contrasts of interest: (a) *moral words - moral non-words*; (b) *non-moral words - non-moral non-words;* (c) *moral words - non-moral words*; and (d) *moral non-words - non-moral non-words*. (Unsurprisingly, the last contrast revealed that the moral non-words and the non-moral non-words showed essentially identical ERP activity).

The contrasts comparing words and non-words within the moral and non-moral stimulus sets were then averaged, leaving two contrasts of interest: (a) *words - non-words*; and (b) *moral words - non-moral words*, which isolate the ERP activity elicited by correctly identifying a stimulus as a word, and by identifying a word as moral, respectively. We reasoned that

participants would first need to determine if a stimulus was a word before they could then determine if the word was moral or not. The grand-average difference waves for these two contrasts were plotted.

To minimize researchers-degrees-of-freedom, we selected the windows for each individual ERP component after looking at the overall data averaged over *all* conditions (see Luck & Gaspelin, 2016). We calculated the largest voltages for each component first, collapsed across all conditions, then selected electrodes where overall activity was greatest. Then we searched for our hypothesized differences across conditions. We chose the electrodes with the largest voltage for each of these components first, collapsed across all conditions, and then looked at differences between conditions. We then separated the data into conditions to compare inside each component. Three later components of interest were identified (EPN, 200-350 ms; P3, 350-600 ms; and LPP, 600-800 ms), and the scalp topographies of the average voltage across all electrodes within these windows were plotted. Based on these topographies we chose single electrodes to help visualize the effects for each of the three components of interest, depending on which electrode exhibited the largest voltage difference from baseline in that component. The EPN component and the P3 component were both strongest over electrode Pz. The LPP component was largest over Cz. Lastly, we plotted the grand average ERPs, the topographies of the difference for each of the two contrasts, and the within-subject difference waves for the two contrasts for these electrodes.

Results

Behavioral Results: Moral Pop-out Effect

First, we hypothesized that moral words would be more likely to be seen than matched non-moral words—a phenomenon known as the *moral pop-out effect;* Gantman & Van Bavel;

2014; 2016). Following our previous work suggesting that this effect is greatest under conditions of ambiguity, we decided *a priori* to present stimuli that would be perceptually ambiguous (e.g., due to their font size, color, and presentation duration). Indeed, accuracy overall (including non-words) was close to the threshold for visual awareness (M = 75%, SE = .4%), roughly halfway between chance responding and perfect accuracy. Replicating our earlier work, moral words (M = 78.5%, SE = .5%) were detected more frequently than non-moral words (M = 75.7%, SE = .5%; B = .158, SE = .043, 95% CI [.073, .242], $p < .001$, z = 3.66); see Figure 2; Gantman & Van Bavel, 2014; 2015) which were both detected more frequently than non-words, which were created by scrambling moral words (M = 71.0%, SE = .6%) versus non-words created by scrambling non-moral words M = 72.8%, SE = .6%).

Overall, participants responded faster to words (M = 553ms, SE = 2ms) versus non-words (M = 611ms, SE = 2ms; (B = -.06, SE = .007, 95% CI [-.07, -.05], $p < .001$, z = -9.14).; trials with no responses have been trimmed from this analysis). Moreover, within word trials, participants responded marginally faster to moral words (M = 550, SE = 3ms) versus non-moral words (M = 556ms, SE = 3ms;), though this difference was not statistically significant (B = -.005, SE = .003, 95% CI [-.010, .001], p = .082, z = -1.74).[2]

[2] We do not consistently find differences in reaction times when stimulus durations are short (~40 ms), however, when stimulus durations are longer (and accuracy gets closer to ceiling) we would predict faster reaction times to moral (vs. non-moral words). Indeed, we have found this pattern in prior work where stimulus durations ranged from 10-100 milliseconds.

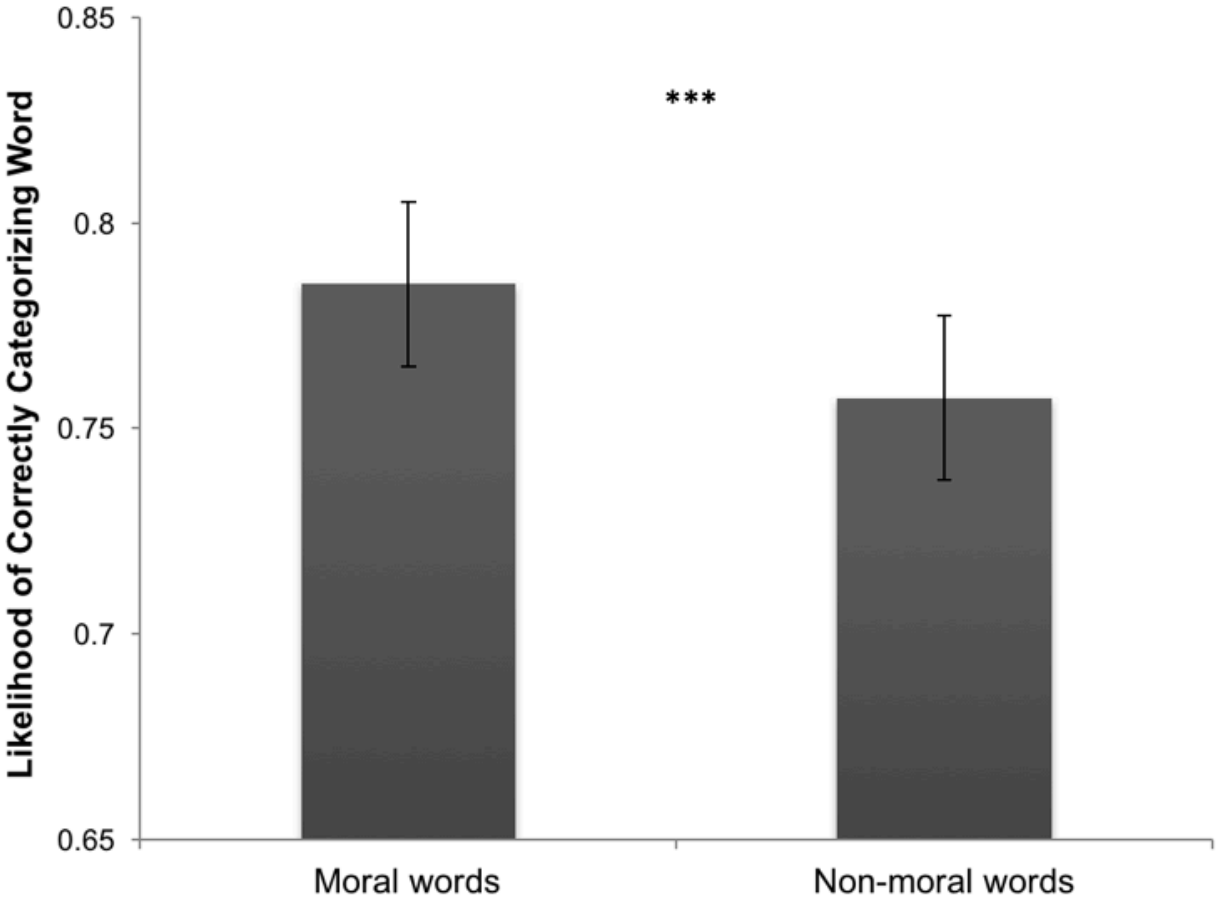

Figure 2. *Predicted means of words correctly identified as words, as estimated from GEE model. Replicating previous research, moral words (79%) were more frequently identified as words than non-moral words (76%). Error bars represent 95% confidence intervals.*

EEG Results: Neural Correlates of Distinguishing Words from Non-words

We first sought to test how early simple distinctions could be observed between words and non-words. When analyzing the ERP results, we included all trials surviving artifact rejection (the moral pop-out effect remains statistically significant in this subsample of trials, B = .16, SE = .05, 95% CI [.06, .26], p = .002, z = 3.11) and then included only trials where

participants correctly identified the target as a word or a non-word.[3] The ERP signature of word discrimination can be observed in Figure 3A, which plots the averaged stimulus-locked activity for each of the four conditions at electrode Pz. A large positive deflection (max ~ 3uV) was observed when participants were presented with words compared to non-words. The difference in Pz voltage between words and non-words was evident for the P2 window (200-250 ms; B = .60, SE = .10, 95% CI [.41, .80], $p < .001$, $z = 6.02$), the N2 window (250-350 ms; $B = 1.23$, SE = .12, 95% CI [.99, 1.45], $p < .001$, $z = 10.38$) and the P3 window (350-600ms; $B = 1.50$, SE = .13, 95% CI [1.26, 1.75], $p < .001$, $z = 12.00$), but not at the LPP time window (600-800ms; B = -.14, SE = .12, 95% CI [-.38, .11], $p = .27$, z = -1.10). As shown in Figure 3C, words and non-words began to be distinguished roughly 200 ms after stimulus presentation in the P2 component time window and continued for roughly 400 ms.

Plotted at Pz, the later difference appeared to take the form of two components (Figure 3C). The topography of this difference over the scalp was fronto-central initial and became more left-posterior over time (Figure 3B). In sum, words and non-words were distinguished as early as 200 ms after word onset, potentially at the level of the visual word form area in the bottom-up ventral visual pathway (McCandliss, Cohen, & Dehaene, 2003).

[3] We chose to analyze only the correct trials, because on a trial with no response, we do not know *why* participants missed the trial. It is possible that they were paying attention and simply failed to see the letter string, but it's also possible that they were distracted in some way during the trial. We reasoned that we would still be able to see an early effect of moral relevance if we could see differences in brain activity attuned to word recognition unfolding at the same time as differences in brain activity attuned to moral relevance.

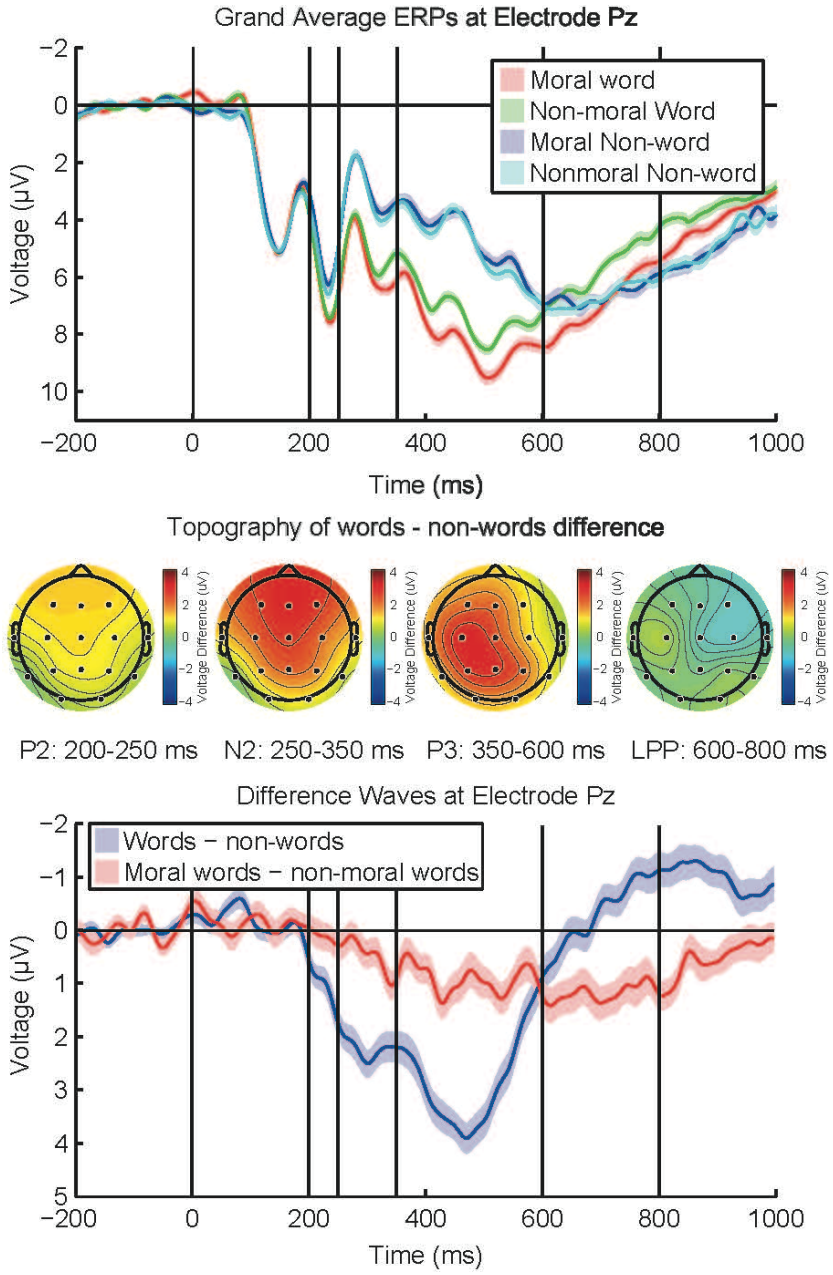

Figure 3. *Top panel) Grand-average ERPs of the average post-stimulus activity at electrode Pz for all four conditions separately. Middle panel) Scalp topographies of the magnitude of the word - non-word difference, averaged across moral and non-moral words and non-words. The difference in activity due to the stimulus being a word or a non-word is largest over fronto-central sites for the N250 component, largest over left-parietal sites for the P3 component, and not observed for the LPP component. Bottom panel) Difference wave of the activity difference between words and non-words (red) and of the difference between moral and non-moral words (blue) at electrode Pz. The difference between words and non-words reliably spans the N250 and P3 component time ranges. Shading represents within subject standard error (i.e., adjusting for between subject differences (see Mason & Loftus, 1995).*

EEG Results: Neural Correlates of Distinguishing Moral from Non-moral Words

Next, we investigated our central research question: When in the perceptual processing stream are moral words differentiated from non-moral words? If moral words are differentiated prior to the onset of the word/non-word distinction, this would provide evidence that the effect is due to differences in early visual processing. If, however, moral words are prioritized immediately following the word vs. non-word effects, this would represent evidence that the content of moral words may be extracted after basic word processing. A more subtle positive deflection (max ~ 1uV) was observed when moral words were presented, compared to non-moral words (Figure 4A). There was no observable difference in Cz voltage between moral and non-moral words at P2 (200-250 ms; B = .07, SE = .10, 95% CI [-.13, .27], p = .521, z = 0.64) but a difference emerged shortly thereafter at N2 (250-350 ms; B = .30, SE = .11, 95% CI [.08, .51], p = .006, z = 2.73), and grew larger at the P3 window (350-600 ms; B = .48, SE = .13, 95% CI [.24, .73], $p < .001$, z = 3.85), and the LPP time window (600-800 ms; B = .46, SE = .15, 95% CI [.17, .75], p = .002, z = 3.11). Moral words and non-moral words were distinguishable in the neural data as early as ~300 ms after word presentation.

Plotted at Cz, this difference seemed to take the form of a single, long-lasting positive deflection from ~300-850 ms (Figure 4C). Similar scalp topographies within the P3 component and LPP component windows showing a posterior-central maxima similar to the canonical P3 topography support this interpretation (Figure 4B). Thus, moral words begin to receive different visual processing roughly 100 ms after words and non-word processing begin to differ. Further, the difference appears similar to a classic P3 component in topography and morphology, indicating the effect of morality on word processing may reflect morally-relevant words receiving preferential access to conscious awareness, via selective attentional enhancement

(Dehaene et al., 2006; Kouider et al., 2013) and that early visual encoding of the word/non-word distinction may have been unaffected by its moral content.

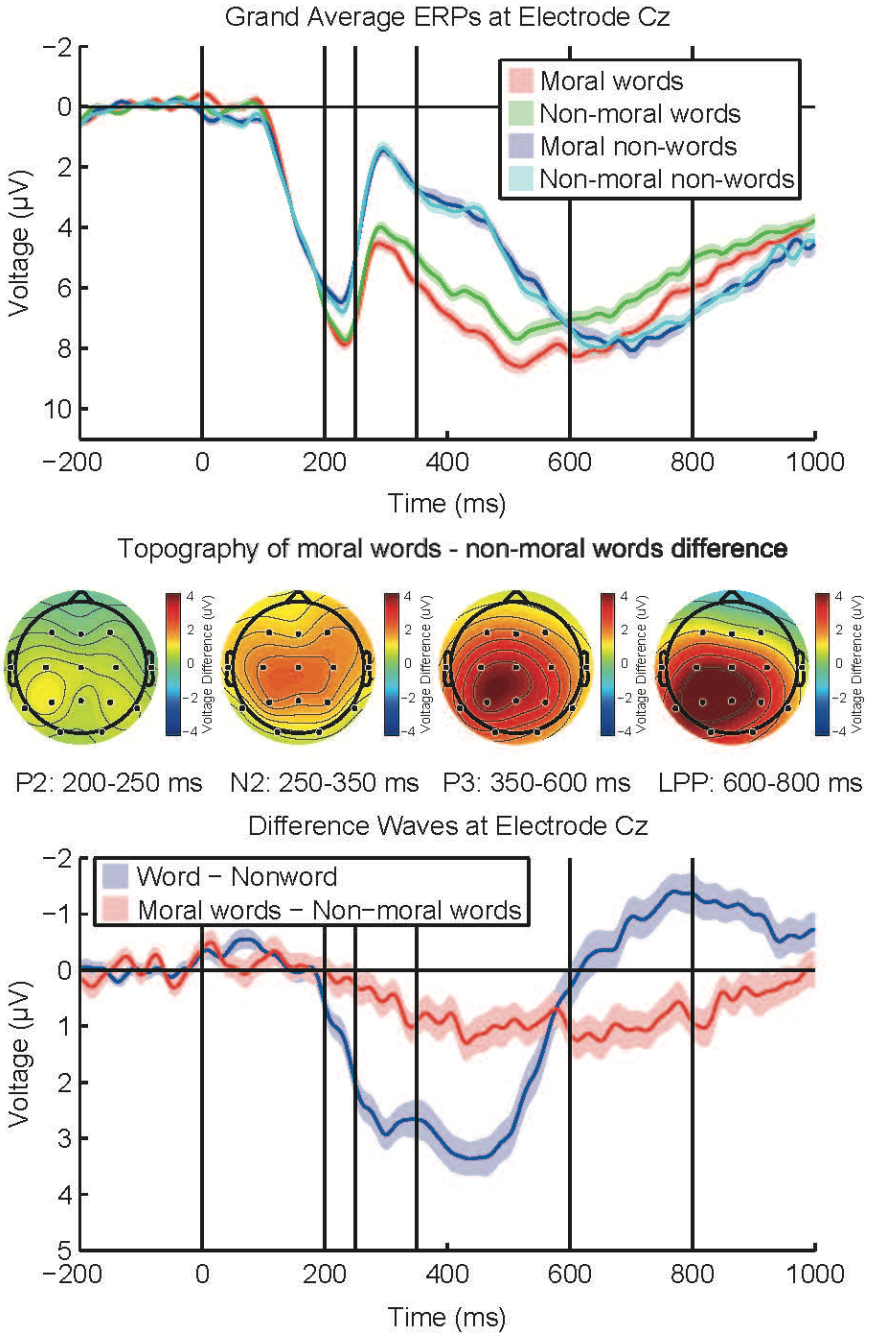

Figure 4. *Top panel) Grand-average ERPs of the average post-stimulus activity at electrode Cz for all four conditions separately. Middle panel) Scalp topographies of the magnitude of the moral - non-moral difference, for words only. The difference in activity due to the stimulus being a moral or a non-moral word is largest over left-parietal sites for the P3 component and for the LPP component, and not observed for the N250 component. Bottom panel) Difference wave of*

the activity difference between words and non-words (red) and of the difference between moral and non-moral words (blue) at electrode Cz. The difference between moral and non-moral words is significant at the P3 and LPP component time ranges. Shading represents within subject standard error (i.e., adjusting for between subject differences (see Mason & Loftus, 1995).

Exploratory analyses of arousal and valence. It is critical to distinguish the present effects of moral relevance from potential effects of valence and arousal, which are fundamental dimensions for categorizing emotionally significant stimuli (Russell & Barrett, 1999). Indeed, prior evidence suggests that stimuli that are emotionally arousing preferentially recruit attention and show variation at LPP (Cuthbert, Schupp, Bradley, Birbaumer, & Lang, 2000; Lang & Davis, 2006; Olofsson, Nordin, Sequeira, & Polich, 2008). Similarly, there is a long history of effects of negativity dominance in psychology (e.g., Rozin & Royzman, 2001) and our earlier pilot data suggest that moral words are perceived as more negative than non-moral words (Gantman & Van Bavel, 2014). Accordingly, it is critical to test whether word category (moral vs. non-moral) explains variation in P3 activity, or if the effects describe above simply reflect the influence of arousal and/or valence.

To test these questions, we utilized human-coded arousal and valence ratings for moral and non-moral words in our experiment from a database of 13,915 word ratings (the 'extended ANEW' set; Warriner, Kuperman, & Brysbaert, 2013). We were able to find valence and arousal ratings for 225 of our 250 words (both moral and non-moral; a list of words without valence and arousal ratings are available on OSF: https://osf.io/5jmze/). We then conducted the same analyses detailed above, but replacing arousal ratings with moral vs. non-moral word coding. We tested the effect of arousal at all four timepoints specified above (P2, N2, P3, and LPP). There was no observable effect of arousal at any timepoint except at the P3 window (350-600 ms; B = .38, SE = .12, 95% CI [.10, .58], p = .006, z = 2.76), such that more arousing words were associated with greater P3 amplitudes. However when word category (moral vs. non-moral

coding) and arousal (continuous ratings) were modeled simultaneously, they both separately affect voltage at the P3 window, with moral (versus non-moral) words (B = .43, SE = .14, 95% CI [.15, .70], p = .002, z = 3.04) showing a qualitatively larger effect than highly arousing words (B = .27, SE = .12, 95% CI [.04, .51], p = .02, z = 2.27).

We conducted the same analysis with valence ratings, and found observable effects of valence at both the P3 window (350-600 ms; B = -.20, SE = .06, 95% CI [-.07, -.33], p = .002, z = 3.09) and LPP window (600-800 ms; B = -.18, SE = .09, 95% CI [-.01, -.35], p = .04, z = 2.11) such that more negative words were associated with greater amplitudes during both windows. However when word category (moral vs. non-moral) and valence (continuous ratings) were modeled simultaneously, they both separately affected voltage at the P3 window, with moral words (B = .40, SE = .15, 95% CI [.11, .70], p = .006, z = 2.74) showing a qualitatively larger effect than more negative words (B = -.15, SE = .07, 95% CI [-.28, -.02], p = .03, z = 2.23). During the later LPP window, when we model both valence and morality, we find that only the effect of word category (moral vs. non-moral) remains (B = .43, SE = .16, 95% CI [.12, .74], p = .007, z = 2.71). These results suggest that our theoretically driven categorization of moral vs. non-moral words is able to explain unique variance in brain activity that is separate from variance explained by arousal or valence.

Exploratory repetition analysis. It has been suggested that the moral pop-out effect can be explained by differences in memory rather than differences in perception (Firestone & Scholl, 2014). The authors of this critique proposed that evidence of repetition priming would be suggestive that memory is playing a key role in the frequent correct categorization of moral (vs. non-moral) words. From this perspective, there would be an advantage for a given moral word if it had been preceded by a moral word but not by a non-moral one. It is worth noting then, that

whether a trial was a repeated trial (e.g., a moral word preceded by a moral word or a non-moral word preceded by a non-moral word) has no effect on differences in ERP activity, (p = .69). This suggests that changes in the P3 window reported here are not reliably tracking potential effects of repetition. Moreover, further exploratory analyses revealed no effect of repetition for moral words only or for non-moral words only (all ps > .05). It is unlikely that differences in P3 activity reported here can be explained by differences in repetition, and thereby, do not strongly suggest that differences in memory are solely responsible for the moral pop-out effect. It is worth noting that this is not to say that repetition does not or cannot affect overall accuracy in our lexical decision task. We present these analyses simply to show that repetition does not explain away the effect that moral relevance has on whether a word reaches conscious awareness.

Discussion

This paper examined when morality is prioritized in perceptual awareness. We used a combination of behavioral and neuroscientific methods to determine when in the perceptual processing stream moral words become more likely to be seen than non-moral words. First, we successfully replicated the moral pop-out effect, such that participants were more likely to correctly identify moral words compared to matched non-moral words (Gantman & Van Bavel, 2014; 2016). Second, we found that moral words were differentiated from non-moral words as early as 300 milliseconds after presentation—which was roughly one hundred milliseconds after people had encoded the difference between words and non-words. Finally, we found that differences in brain activity for moral vs. non-moral words cannot be explained by differences in the valence of or arousal related to the words. These results most strongly align with Hypothesis 2 stated in the Introduction: morality affects what reaches perceptual awareness. Specifically, moral words were prioritized over non-moral words later in perceptual processing, reflecting a

post-perceptual cognitive process (Squires, Donchin, Herning, & McCarthy, 1977; Polich, 2012), which prioritizes moral (vs. non-moral) content to reach conscious awareness (Dehaene et al., 2006; Kouider et al., 2013, though for a dispute regarding this interpretation of P3 activity, see Cohen, Ortego, Kyroudis, & Pitts, 2020, Jan 16).

Time-course of Moral Word Identification

Overall, the neural data we observed suggest two independent effects of our stimuli on subsequent ERP amplitudes. Participants begin to differentiate words from non-words as early as 200 milliseconds after stimulus presentation and primarily over frontal recording sites. Very soon after participants encoded these words, the difference in activity between words and non-words transitioned to left-posterior recording sites, where moral and non-moral words began to be discriminated by 300 milliseconds post-presentation. Thus, the shift in information processing was observed in time as well as space—with activity moving from frontal to posterior cortex as moral words were encoded and prioritized in perceptual processing. Of course, the spatial resolution of EEG is relatively coarse and limited. With that in mind, we encourage future work using functional magnetic resonance imaging to achieve a more precise understanding of the neural correlates of the moral pop-out effect in particular, and moral perception in general.

It is important to note that moral content remained privileged in information processing as higher order cognition came online. By 600 milliseconds, the ERP activity could no longer discriminate whether a word or a non-word had been presented. However, there was a consistent difference in ERP activity between moral and non-moral words until approximately 850 milliseconds after the stimuli were presented suggesting a sustained preference for moral content. We tentatively interpret this finding as a threshold effect: the emergence of differential P3 activity for moral vs. non-moral words suggests that moral words may receive and maintain a

preferential gateway into conscious awareness. As such, it appears that the moral pop-out effect is likely not a perceptual pop-out, but perhaps a pop-in, to awareness.

Moreover, we note that our exploratory analyses rule out the possibility that the increase in P3 (as well as LPP) magnitude for moral vs. non-moral words can be attributed to factors confounded with morality. While previous work demonstrates the sensitivity of these waveforms to arousing stimuli (e.g., Olofsson et al., 2008), these results suggest that the moral content of words presented in our task captured awareness over and above both arousal and valence. Finally, despite the relatively mixed ERP literature on moral cognition, we note that these results dovetail with previous findings indicating that morality exerts its influence on processing about 300 ms after the presentation of lexical information (Yang et al., 2014; Yang et al., 2017). We note that other research on the time-course of moral responses using photographic images as stimuli has observed morality-specific effects in earlier time windows (e.g., N1, Yoder & Decety, 2014; Gui et al., 2016; see also Decety & Cacioppo, 2012), while work examining the temporal dynamics of moral judgments using vignettes observes effects of morality in later time windows (e.g., LPP, Leuthold et al., 2015). Future work should directly assess how differences in stimulus format influence the time-course of moral perception.

Ultimately, we chose to focus on the presentation of morally relevant words for a few key reasons. First, moral words are more likely to be seen than non-moral words—a phenomenon termed the *moral pop-out effect*. Critically, the moral pop-out effect only occurs when letter strings are presented ambiguously—near the threshold for visual awareness (Gantman & Van Bavel, 2014). We recognize that the differences in accuracy for moral (79%) vs. non-moral words (76%) is not particularly large. However, there is reason to think it is both theoretically and practically important. Although confidence intervals are the best measure of effect size for

GEE analyses, we can utilize the z-score to get a rough estimate, $r^2 = .27$, a moderate effect in Psychology (Ferguson, 2009). Theoretically, this experiment expands on this prior work by clarifying the underlying neural processes that precede moral word identification. Practically, people consume and share enormous volumes of moral language—especially on social media (e.g., Twitter). By one estimate, people scroll through 300 feet of social media content per day. An effect of this size could accumulate over such a large volume of content. People are now more likely to experience moral outrage from learning about an event online than in real life (Crockett, 2017) and are more likely to share messages (e.g., retweet a Tweet) that contain moral-emotional (vs. neutral) content in political conservations (Brady et al, 2017). The perceptual salience of moral words can help us understand why. Tweets that contain moral language are more likely to be seen, and the degree to which they capture our attention is related to real online sharing behavior (by a completely separate group of users) on Twitter (Brady, Gantman, & Van Bavel, 2019). As such, understanding how and when moral word processing affects perception is timely and consequential. That said, future work should examine whether these processes apply to other, more ecologically valid, moral stimuli in addition to moral words presented alone, as we did here.

Another important caveat to this work is our sample. To study moral perception using words, we were limited to English-speaking students. Unfortunately, we do not yet know if the results generalize to other populations. We welcome and encourage future work on the moral pop-out effect with a variety of samples, and further suggest that words that are deemed morally relevant by participants may vary with different samples and should be tailored to different populations. This approach would not only examine the generality of our findings, but might identify critical boundary conditions that elucidate new aspects of human morality and vision.

Given that morality is culturally constructed, we think this is an important and fruitful avenue for future research.

Conclusion

The present research utilized electroencephalography to examine the time-course of moral perception. Previous research has suggested that moral words are more likely to be seen than non-moral words, and speculated that this effect occurs at the level of visual experience rather than memory or response preparation. To disentangle these possibilities, we utilized a temporally sensitive measurement of brain activity in order to observe *when* in the perceptual processing stream moral words bias brain activity. We observed that moral stimuli were distinguished from non-moral stimuli within a few hundred milliseconds (~300 ms after stimulus onset)—immediately after people had encoded whether the letter string was a word or not. This work replicates and clarifies previous research on moral perception, while also extending our understanding of the social factors that may heighten conscious awareness. This work suggests that moral content is more likely to be seen because it lowers the threshold for being broadcasted to conscious awareness.

References

Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact of gossip. *Science*, *332*, 1446-1448.

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences, 106*, 1672-1677.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*, 551–565.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*, 7313-7318.

Brady, W.J., Gantman, A.P. & Van Bavel, J.J. (2019). *Journal of Experimental Psychology: General.* Advance online publication. *https://doi.org/10.1037/xge0000673*

Callan, M. J., Ferguson, H. J., & Bindemann, M. (2013). Eye movements to audiovisual scenes reveal expectations of a just world. *Journal of Experimental Psychology: General, 142*, 34-40.

Cohen M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020, January 16). Distinguihsing the neutral correlates of perceptual awareness and post-perceptual processing. *BioRxiv preprint.* doi: http://dx.doi.org/10.1101/2020.01.15.908400.

Cunningham, W. A., Packer, D. J., Kesek, A., & Van Bavel, J. J. (2009). Implicit measures of attitudes: A physiological approach. In R. E. Petty, R. H. Fazio & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 485–512). New York: Psychology Press.

Cunningham, W.A., Van Bavel, J.J., Arbuckle, N.L., Packer, D.J., & Waggoner, A.S. (2012). Rapid social perception is flexible: Approach and avoidance motivational states shape P100 responses to other-race faces. *Frontiers in Human Neuroscience, 6,* 140.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769.

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P. J. (2000). Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biological psychology*, *52*, 95-111.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990 present. Available online at http://corpus.byu.edu/coca.

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of neurophysiology*, *108*(11), 3068-3072.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences, 10*, 204-211.

Everett, J. A. C., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Science, 24,* 35-45.

Firestone, C., & Scholl, B. J. (2015a). Enhanced visual awareness for morality and pajamas? Perception vs. memory in 'top-down'effects. *Cognition, 136*, 409-416.

Firestone, C., & Scholl, B. J. (2015b). 'Moral perception' reflects neither morality nor perception.

Trends in Cognitive Sciences, 20, 75-76.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for 'topdown' effects. B*ehavioral & Brain Sciences*, e229, 1-77.

Friedman, D., & Johnson, R. E. (2000). Event-related potential (ERP) studies of memory encoding and retrieval: A selective review. *Microscopy Research and Technique, 51,* 6 28.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, *45*, 152-170.

Gan, T., Lu, X., Li, W., Gui, D., Tang, H., Mai, X., ... & Luo, Y. J. (2016). Temporal dynamics of the integration of intention and outcome in harmful and helpful moral judgment. *Frontiers in psychology*, *6*, 2022.

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition, 132,* 22-29.

Gantman, A. P., & Van Bavel, J. J. (2015a). Moral perception. *Trends in Cognitive Sciences, 19*, 631-633.

Gantman, A. P., Van Bavel, J. J. (2015b). Letter: See for yourself: Perception is attuned to morality. *Trends in Cognitive Sciences, 20,* 76-77.

Gantman, A. P., Van Bavel, J. J. (2015c). Commentary on Firestone and Scholl: Behavior is multiply determined and perception is multiply defined. *Behavioral Brain Sciences*.

Gantman, A. P., & Van Bavel, J. J. (2016). Exposure to justice diminishes moral perception. *Journal of Experimental Psychology: General, 145,* 1728-1739.

Gilden, L., Vaughan Jr, H. G., & Costa, L. D. (1966). Summated human EEG potentials with voluntary movement. *Electroencephalography and clinical Neurophysiology*, *20*, 433-438.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, *6*(12), 517-523.

Gray, H. M., Ambady, N., Lowenthal, W. T., & Deldin, P. (2004). P300 as an index of attention to self-relevant stimuli. *Journal of experimental social psychology*, *40*(2), 216-224.

Gui, D. Y., Gan, T., & Liu, C. (2016). Neural evidence for moral intuition and the temporal dynamics of interactions between emotional processes and moral cognition. *Social neuroscience*, *11*, 380-394.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and psychological bases of ideology and system justification* (pp. 371– 401). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195320916 .003.015

Haidt, J. (2008). Morality. *Perspectives on Psychological Science, 3,* 65-72.

Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology, 36*, 660–676.

Ito, T. A., Thompson, E., & Cacioppo, J. T. (2004). Tracking the timecourse of social perception: the effects of racial cues on event-related brain potentials. *Personality and Social Psychology Bulletin, 3*0, 1267–1280.

Johnston, V. S., Miller, D. R., & Burleson, M. H. (1986). Multiple P3s to emotional stimuli and their theoretical significance. *Psychophysiology*, *23*, 684-694.

Kohlberg, L. (1979). *The meaning and measurement of moral development*. Clark University Press.

Kouider, S., Stahlhut, C., Gelskov, S. V., Barbosa, L. S., Dutat, M., De Gardelle, V., ... &

Dehaene-Lambertz, G. (2013). A neural marker of perceptual consciousness in infants. *Science, 340*, 376-380.

Lang, P. J., & Davis, M. (2006). Emotion, motivation, and the brain: reflex foundations in animal and human research. *Progress in Brain Research*, *156*, 3-29.

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Social Cognitive and Affective Neuroscience*, *10*, 1021-1029.

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Luck, S. J., & Gaspelin, N. (2016) How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, 54, 146-157.

Luck, S. J., & Kappenman, E. S. (Eds.). (2011). The Oxford handbook of event-related potential components. Oxford university press.

Luo, Y., Shen, W., Zhang, Y., Feng, T. Y., Huang, H., & Li, H. (2013). Core disgust and moral disgust are related to distinct spatiotemporal patterns of neural processing: An event-related potential study. *Biological Psychology*, *9*, 242-248.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Eeview*, *1*, 476-490.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences, 7,* 293-299.

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: an integrative review of ERP findings. *Biological psychology*, *77*, 247-265.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in cognitive science*, 2(3), 511-527.

Polich, J. (2012). Neuropsychology of P300. *Oxford handbook of event-related potential components*, 159-188.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296-320.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, *76*, 805-819.

Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., & Palomba, D. (2012). Temporal dynamics of cognitive-emotional interplay in moral decision-making. Journal of Cognitive Neuroscience, 24(4), 1018–1029.

Schupp, H. T., Cuthbert, B. N., Bradley, M. M., Cacioppo, J. T., Ito, T., & Lang, P. J. (2000). Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology*, *37*, 257-261.

Smith, N. K., Cacioppo, J. T., Larsen, J. T., & Chartrand, T. L. (2003). May I have your attention please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia, 41*, 171–183.

Squires, N. K., Donchin, E., & Squires, K. C. (1977). Bisensory stimulation: Inferring decision related processes from the P300 component. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 299-315.

Strayer, D. L., & Johnston, W. A. (2000). Novel popout is an attention-based phenomenon: An ERP analysis. *Perception & psychophysics*, *62*, 459-470.

Tamietto, M., & De Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience*, *11*, 697-709.

Yang, Q., Li, A., Xiao, X., Zhang, Y., & Tian, X. (2014). Dissociation between morality and disgust: An event-related potential study. *International Journal of Psychophysiology*, *94*, 84-91.

Yang, Q., Luo, C., & Zhang, Y. (2017). Individual differences in the early recognition of moral information in lexical processing: An event-related potential study. *Scientific reports*, *7*, 1475.

Yoder, K. J., & Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. *Neuropsychologia*, *60*, 39-45.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233-235.

Zhu, R., Wu, H., Xu, Z., Tang, H., Shen, X., Mai, X., & Liu, C. (2019). Early distinction between shame and guilt processing in an interpersonal context. *Social neuroscience*, 14(1), 53-66.