Less Than Reckless:
Assessing the Role of Consciousness in the Moral Appraisal of Risky Action

A thesis presented by

Ilai Gavish

to
the Faculty of the Committee on Degrees in Neuroscience and Philosophy
in partial fulfillment of the requirements
for the degree with honors
of Bachelor of Arts
and Certificate in Mind, Brain, & Behavior

Harvard University
Cambridge, Massachusetts
March 2023

Neuroscience Concentration

Division of Life Sciences

Harvard University

The Harvard College Honor Code

Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

*Signature:*____________Ilai Gavish____

Acknowledgements:

In the abstract, a thesis is a collection of ideas: hypotheses, variables, hypotheticals, rebuttals, and the like. In the concrete, it is a collection of people's efforts, of which the author's work forms only a part. To the countless people who shaped this thesis and its author, I give my thanks. In particular, I would like to thank the following individuals who made this thesis possible:

To Uri Maoz, for welcoming me to the Brain Institute and encouraging me to pursue the ideas that interest me most.

To Tomáš Dominik, Emma Chen, and Alison Oliver, for your ever-reliable and immeasurable help. This thesis would be a complete mess, and a far less fun endeavor, without you.

To Gabriel Kreiman, for your continuous support and flexibility from the beginning.

To Zoë Johnson King, for your attentive guidance and clear-headed insights, without which my philosophical ideas would be a fraction as developed. I could not ask for a better advisor.

To Ima, Abba, Yarden, Yotam, Einat, Tom, and Shelby, for the knowledge that no matter how things turn out, I will always have the best family in the world. I love you.

To Jacqueline Gong, Wendy Li, and Jacob Jampel, for always rooting for me along the way, even when you had no idea what I was talking about.

To Logan Dick, Paul Apostolicas, Cameron Hamby, Graydon Moorhead, and Ben Zhang, for tolerating my endless thesis talk and bottom-of-the-barrel jokes, and for being wonderful blockmates. Semper Cor!

To Cecilia Zhou, Juliet Isselbacher, Annie Miall, Chloe Shawah, Isabel Diersen, Daniel Abdulah, and so many others whom I wish to name, for making my time at Harvard truly magical.

List of Contributions:

This project was conceived by Ilai Gavish with the guidance of Uri Maoz, Gabriel Kreiman, and Gideon Yaffe. The metacontrast experiment was designed by Ilai Gavish with help from Tomáš Dominik, Uri Maoz, Gabriel Kreiman, and Liad Mudrik. Electrophysiological recordings were done jointly by Ilai Gavish, Tomáš Dominik, Emma Chen, and Alison Oliver. Data and results were interpreted by Ilai Gavish with assistance and guidance from both Tomáš Dominik and Uri Maoz. The diagram in **Figure 1** was generated by Luck (2014), and the diagram in **Figure 2** was generated by Mastropasqua & Turatto (2015). The philosophical concepts were analyzed by Ilai Gavish with guidance from Zoë A. Johnson King.

Abstract:

The law typically defines criminal recklessness as having *conscious awareness* of an unjustifiable risk of harm and choosing to act despite this risk. This thesis investigates the validity of this requirement of conscious awareness using the methods of both neuroscience and philosophy. First, I conducted an electroencephalography (EEG) experiment in which subjects were presented with a binary choice wherein one of the options was sometimes preceded by a stimulus signaling risk of harm to a future participant. This risk-stimulus was presented either consciously or subliminally using a metacontrast masking paradigm. In some analyses, electrical activity at the midline central (Cz) electrode showed a significantly greater post-choice P300 amplitude for risky trials than for trials without a signal of risk, and there was significant interaction with the conscious/unconscious presentation of the stimulus as well as with the strategy employed by the participant. These preliminary results suggest that there is a detectable difference in neural activity between conscious versus unconscious processing of risk. This neuroscientific experiment is supplemented by a broader philosophical discussion of the relationship between consciousness and moral responsibility. I argue that volitionalism, which requires conscious awareness for blameworthiness, prevails over consciousness-optional views on theoretical grounds, and I rebut a set of anti-volitionalist moral intuitions by introducing a distinction between the concepts of responsibility and ownership. In combination, the neuroscience and philosophy research helps to validate both that there is a distinction between conscious and unconscious representation of risk and that this difference is morally meaningful.

Table of Contents:

Title Page 1

Honor Code 2

Acknowledgements..3

List of Contributions 4

Abstract 5

Table of Contents 6

List of Tables and Figures..7

Introduction..8

Methods..18

Results..28

Empirical Discussion 38

Philosophical Discussion 45

Supplementary Materials 72

References..82

List of Tables and Figures:

Figure 1 13

Figure 2 16

Figure 3 19

Figure 4 21

Figure 5 22

Figure 6 29

Figure 7 31

Figure 8 32

Figure 9 34

Figure 10 36

Figure S1 72

Figure S2 73

Table S1 74

Table S2 74

Table S3 75

Table S4 76

Table S5 77

Table S6 79

Table S7 81

Introduction:

I. Moral Responsibility, Consciousness, and Reckless Action

Human beings are moral creatures. Our actions are assessed normatively in terms of good and bad, blameworthy and praiseworthy. However, much of our behavior arises from information and attitudes of which we have no awareness. The brain, the primary cause of action (moral, immoral, and amoral), absorbs and represents far more information than what enters consciousness, and these unconscious processes shape our actions in countless ways (Mudrik & Deouell, 2022). For instance, online advertisements that a consumer does not pay attention to can still influence their purchasing choices (Yoo, 2008), and implicit race and gender biases can lead to unintended discriminatory behavior (Greenwald & Krieger, 2006). Consequently, a challenge for our moral theories, and our legal frameworks, is to clarify what role, if any, consciousness should play in our understanding of moral responsibility.

One moral and legal concept whose relationship to consciousness requires further explanation is that of reckless action. "Recklessness" is a mental state referenced frequently in criminal statutes (see *Mass. Gen. Laws, ch. 265 § 13L* or *10 U.S.C. § 914 - Art. 114* for examples), but whose meaning is often ambiguous and hard to apply. The *Model Penal Code*, on which many American laws are based, identifies recklessness as one of the four hierarchical categories of *mens rea*, or culpable state of mind, along with intention, knowledge, and negligence (Legal Information Institute, n.d.). Recklessness is considered less severe than intention (i.e. acting with the purpose of causing harm) or knowledge (i.e. acting while knowing that you *will* cause harm), but more severe than negligence (i.e. acting while unaware of a risk of harm of which you reasonably should have been aware). Specifically, the *Model Penal Code* states that "[a] person acts recklessly with respect to a material element of an offense when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct" (Greenawalt, 1991). One critical aspect of this definition is that recklessness requires *conscious* disregarding of risk—or what the *Model Penal Code* describes as being in a "state of awareness" and engaging in "conscious risk creation" (Greenawalt, 1991)—as opposed to merely storing probabilistic information unconsciously without having direct awareness that one's actions could be harmful.[1]

If this consciousness requirement is accepted, then individuals who act despite an unjustifiable risk that they recognize only unconsciously may not be considered reckless. Take, for instance, the example of a tricky intersection and the aware versus unaware driver. Suppose there is a person who takes the same route to work every day, such that he can complete the route effortlessly on "autopilot," without having to pay attention to the details of the road. Thanks to his repetitive experience, he has unconsciously picked up on the traffic patterns and particular timing of every street and intersection on his route. Suppose there is one intersection along this route where there is poor visibility for pedestrians, meaning that if the driver turns right immediately after the traffic light changes to green, he risks hitting an oblivious pedestrian. For this reason, the driver has learned to automatically wait a moment before turning right. Now consider two different scenarios. In one scenario, the driver is *explicitly* aware that the intersection is set up in this fashion, and he is conscious of the risk associated with turning quickly after the light changes. If, one day, he is late to work and rushing to the office, and he decides to disregard the risk and turn right immediately at the intersection, hitting a pedestrian as a result, his actions would fall under the standard definition of recklessness.[2] In the second scenario, meanwhile, the driver never becomes

[1] For this discussion, it is important to distinguish between the concepts of recklessness and negligence. While recklessness requires (typically conscious) knowledge of a risk, negligence requires no understanding of risk whatsoever—neither conscious nor unconscious. It only requires that a reasonable person in the situation *should have* been aware of the risk. Thus, questions of negligence can be addressed separately from those of recklessness.

[2] There are, of course, a number of factors in this scenario that could affect the extent to which the driver should be held morally or legally responsible, such as why the intersection is set up in this way in the first place, but as this example is only meant to elucidate the conscious versus unconscious distinction, I will not go into them.

explicitly aware of the mechanics of the intersection. Instead, he has only unconsciously represented this detail, and his automatic driving processes lead him safely through the intersection each time. One day, he is late to work and rushing to the office, and he deviates from his habit of pausing at the intersection; instead he turns quickly, hitting a pedestrian in the process. In this case, although he had implicit knowledge of the risk involved in rushing through the intersection (as evidenced by his previous routine behavior at this intersection) he was never *consciously aware* of the risk. As a result, he would not be considered reckless according to the standard definition.

While this distinction arises from the legal definition of recklessness, it remains to be empirically tested whether there is any detectable neural difference between these mental states that would validate the law's separation of conscious and unconscious processing of risk. As Maoz and Yaffe (2016, 135) explain, “To date, few if any neuroscientific studies have investigated the distinctive nature of conscious awareness of risk, distinguishing its neural basis, and role in decision-making, from tacit, or unconscious representations of probabilistic information.” If no such distinction can be established, it would challenge the notion that the mind operates differently when disregarding consciously versus unconsciously represented risk, threatening the standard moral and legal categorization of reckless action. Although I am not aware of any studies directly examining the difference between conscious risk awareness and unconscious representation of probability, there is precedent for the use of neuroscientific tools to validate and elucidate established distinctions between crucial moral-legal concepts, such as Vilares et al.’s (2017) employment of fMRI and machine learning to assess the division between the mental states of knowledge and recklessness.

Accordingly, in the empirical portion of this thesis, I use electroencephalography (EEG) to test whether conscious and unconscious representations of risk are neurally distinguishable. I conducted an experiment in which participants were repeatedly shown stimuli that signaled either the presence or absence of a risk of harm. I designed the experiment using a subliminal perception paradigm called metacontrast masking, such that the stimuli were apprehended consciously at times and only unconsciously at other times, in random sequence. This paradigm makes it possible to compare brain activity in response to four different types of stimuli: conscious risk-signaling, conscious non-risk-signaling, unconscious risk-signaling, and unconscious non-risk-signaling. I analyzed the resulting EEG data by comparing across conditions the sizes of two different EEG waveforms pertinent to cognitive and moral decision-making tasks. In the following section of the introduction, I will describe these waveforms and explain their relevance. Next, in Section III, I will provide a background on the metacontrast masking technique. Finally, I will introduce the conceptual issues I address in my philosophical discussion, and I will give a broad overview of the objectives of this thesis.

II. Event-Related Potentials of Interest

As EEG primarily offers temporal rather than spatial specificity (Im, 2018), one useful methodology for analyzing EEG data is through the identification and examination of event-related potentials (ERPs). ERPs are changes in voltage caused by brain activity in alignment with specific events or stimuli. These changes in voltage are thought to reflect the combination of postsynaptic potentials generated when a large group of thousands to millions of cortical pyramidal neurons fire simultaneously and in phase in the course of information processing (Sur & Sinha, 2009). Different kinds of sensory, cognitive, and motor events elicit different ERPs, each with a specific time-course and directionality.

Two ERP components of interest to the study of moral decision-making are the P2 and P3 waves. The P2 wave, or P200, is a positive spike in voltage that peaks between 150 and 250 ms after stimulus onset (see **Figure 1**) (Luck, 2014; Chen et al., 2009). The P2 component is known to occur with the highest amplitude in fronto-central regions, and it is associated with a variety of cognitive tasks, including selective attention and feature detection processes (Key et al., 2005). Importantly, it has been shown to reflect the initial evaluation of task-relevant stimuli and the onset of the decision-making process (Lindholm & Koriath, 1985; Chen et al., 2009). The P3 wave, or P300, typically occurs between 250 and 450 ms after stimulus onset (see **Figure 1**) (Luck, 2014; Chen et al., 2009; Sur & Sinha, 2009). It is strongest in central parietal areas (Polich, 2011) and theorized to reflect activity in frontal and temporo-parietal brain structures (Key et al., 2005; Nieuwenhuis et al., 2005; Polich, 2007). It has also been linked to the locus coeruleus–norepinephrine system, a subcortical neuromodulatory nucleus thought to enhance responsivity in the neocortex in reaction to motivationally significant stimuli (Nieuwenhuis et al., 2005). The P3 wave is typically elicited in response to unexpected, emotionally valent, and motivationally salient stimuli, and during the processes of attentional allocation, decision-making, and context updating (Nieuwenhuis et al., 2005; Lindholm & Koriath, 1985; Chen et al., 2009; Donchin & Coles, 1988).

In addition to their associations with general information processing events, the P2 and P3 components have been found to be generated during moral dilemma tasks. For example, when Zhan et al. (2020) tasked participants with choosing whether to administer painful electric shocks, either to themselves or to others, in exchange for monetary reward, they observed a positive wave in EEG activity around 160-260 ms and an extended positivity at 300-450 ms after the presentation of the choice, and positivity was greater both when the trade-offs were higher and when the choice involved shocking a stranger rather than oneself. Moreover, in a game involving monetary offers between players, participants showed a greater P3 in response to fair offers than unfair offers, and more prosocial participants also showed a greater P2 in response to fair offers compared to unfair

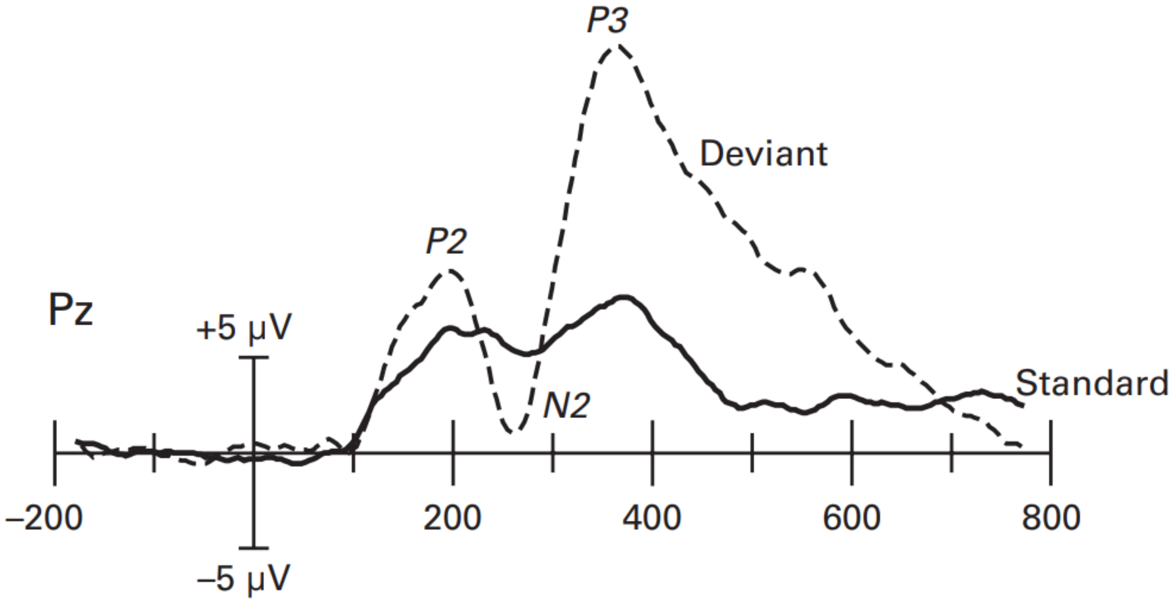


Figure 1 | P2 and P3 Waves. From Luck (2014), an example of P2 and P3 waves elicited at the Pz (midline parietal) electrode site in response to surprising (deviant) stimuli in an "oddball" task. The x-axis is time in ms, time-locked to stimulus onset. The solid line shows the response to standard stimuli for comparison.

offers (Hu & Mai, 2021). Finally, when survivors of the 2008 Sichuan earthquake were presented with a choice of whom to rescue from earthquake debris, they demonstrated a greater P2 and P3 when having to choose between two family members than between two strangers, particularly after being told that an aftershock from the earthquake was going to occur shortly in real life (Chen et al., 2009). All of these findings suggest that the P2 and P3 components reflect an aspect of moral cognition and social-emotional decision-making.

While little research has been done regarding the electroencephalographic correlates of recklessness and risk processing, one study of note is Schmälzle (2008), which identified increased centro-frontal positivity around 300 ms for faces judged as risky compared to non-risky. Consistent with the studies outlined above, I propose that a difference in P2 and P3 generation in scenarios where a risk is apprehended consciously versus unconsciously would suggest that there is a difference in the process of moral cognition when one has conscious awareness of a risk as opposed to a solely implicit representation of the risk.

III. Conscious and Unconscious Presentation of Risk Information

In order to compare neural correlates of conscious and unconscious processing of risk, I designed an experiment in which information about risk of harm to others is signaled either supraliminally or subliminally. There is evidence that participants are able to learn and utilize probabilistic instrumental information conveyed through subliminally presented stimuli (Pessiglione et al., 2008; Mastropasqua & Turatto, 2015). One way in which a stimulus, including a stimulus that signals probability or risk, can be presented and perceived without entering conscious awareness is through a method known as metacontrast masking (Mastropasqua & Turatto, 2015). In metacontrast masking, conscious visibility of a target stimulus is suppressed, or "masked," by the presence of a spatially adjacent stimulus that follows the target stimulus in time

(Breitmeyer et al., 2008). The target stimulus, usually a simple visual shape, is flashed for a brief interval (typically ~10-30 ms), followed by a brief delay and presentation of the "mask" (see **Figure 2)** (Breitmeyer et al., 2008; Mastropasqua & Turatto, 2015; van Gaal et al., 2010). Masking is typically strongest when the mask stimulus is presented 30-80 ms after the onset of the target stimulus (Breitmeyer et al., 2008). Although the target stimulus is not consciously perceived, it can still influence behavior, suggesting that the stimulus information is absorbed non-consciously (Mastropasqua & Turatto, 2015; van Gaal et al., 2010).

IV. An Overview

Although this experiment would assist in confirming the neural and cognitive distinction between conscious and unconscious disregarding of risk, the findings alone would not be sufficient to validate the *moral* distinction between these mental states. For this reason, the thesis will end with a philosophical discussion of the relationship between consciousness and moral responsibility. In this discussion, I will evaluate the arguments made for and against requiring conscious awareness as a necessary condition for moral responsibility, and I will propose an approach that ultimately supports the consciousness requirement and therefore the standard definition of recklessness.

To summarize, this thesis investigates the neural and moral basis for the distinction between conscious recklessness and mere unconscious disregarding of risk. Specifically, I aim to 1) Use metacontrast masking to run an EEG experiment in which participants are faced with a choice that involves consciously presented risk of harm, subliminally presented risk of harm, or no risk; 2) Analyze the behavioral data from the experiment to confirm its conceptual validity; 3) Analyze the resulting EEG data, particularly differences in the P2 and P3 waveforms, using a

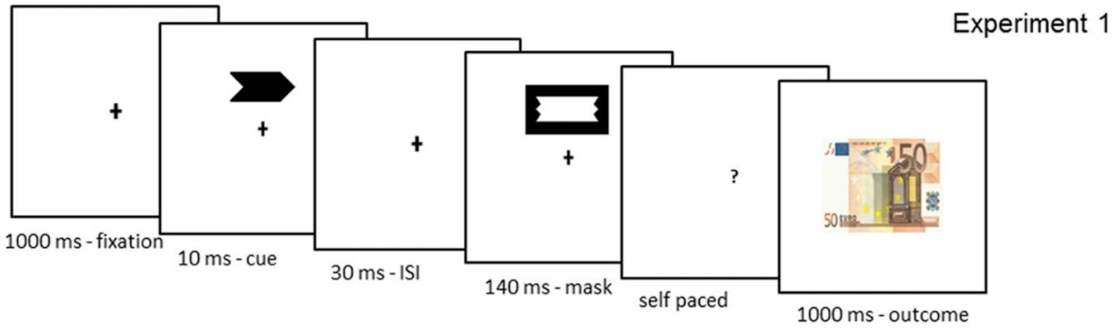


Figure 2 | Metacontrast Masking. From Mastropasqua & Turatto (2015), an example of the metacontrast masking technique. The target stimulus is presented for 10 ms, followed by an interstimulus interval of 30 ms and the presentation of a spatially surrounding mask stimulus for 140 ms. In this specific experiment, the participant is then presented with a choice whose outcome can be favorable (represented by the Euro note) or unfavorable, depending on the masked cue.

repeated measures ANOVA; and 4) Provide a philosophical argument for making conscious awareness a necessary condition for moral responsibility.

Methods:

I. Participants

Participants were students and staff recruited from the Chapman University community in Orange, California. In total, twelve subjects participated in the experiment. Three participants were excluded as their behavioral data indicated they misunderstood the task. Of the remaining nine participants, six were female, two were male, and one was non-binary. The ages of the participants ranged from 20 to 30 with an average age of 21.7. All participants were right-handed and reported having 20/20 vision or vision corrected to 20/20.

II. Stimuli & Procedure

To investigate the neural activity associated with conscious and unconscious representation of risk, I used a novel paradigm in which information about risk of harm to others was presented either supraliminally or subliminally through metacontrast masking. In this task, participants were repeatedly presented with a forced binary choice wherein choosing one of the options led to a monetary reward. Before each choice, participants were presented with a stimulus that was either the shape of a circle or a star (see **Figure 3**). The circle indicated that there was no risk associated with either option of the forced choice, while the star indicated that selecting the reward option would carry the risk of triggering a financial penalty for the next participant in the study. Participants were thus faced with a moral dilemma whenever the risk-associated star was presented. Importantly, using changes in interstimulus interval length and stimulus opacity level, both the circle and the star varied in whether they were presented supraliminally ("unmasked") or subliminally ("masked"), such that the morally-relevant risk information was processed either explicitly or implicitly. In this way, the paradigm allowed for the comparison of EEG activity across the two dimensions of risky versus non-risky and conscious versus non-conscious.

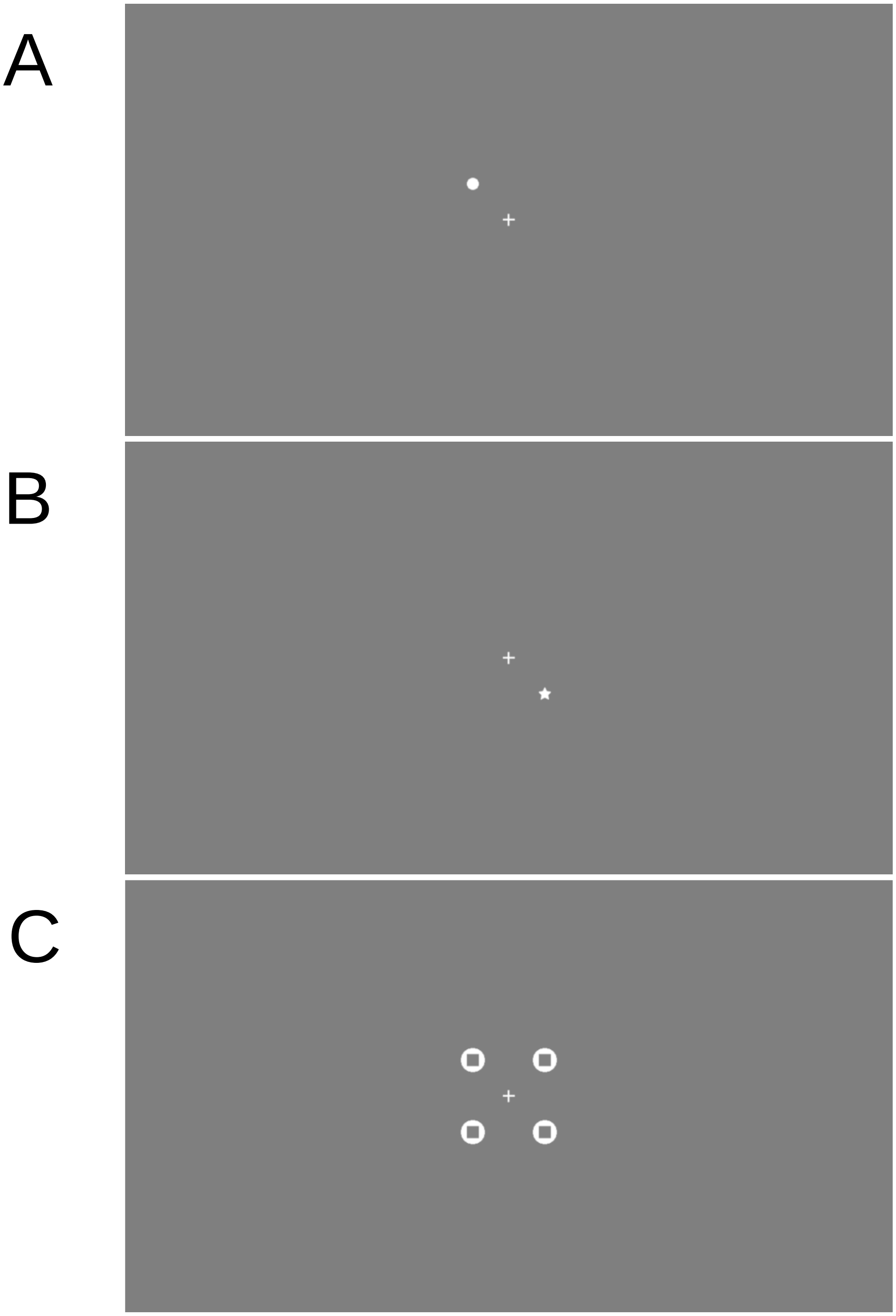


Figure 3 | Experimental Stimuli. The two stimuli shapes used in the experiment were a circle (**A**) and a star (**B**). Each trial, the stimulus was presented for 28 ms in one of four locations around the fixation cross: top left (**A**), top right, bottom left, or bottom right (**B**). After an interstimulus interval of either 28 ms or 280 ms, the mask stimulus (**C**) was presented for 28 ms, with one ring around each of the four possible stimulus locations.

The experiment was coded using the PsychoPy software package (Peirce et al., 2019). The two stimuli shapes used were a white circle and a white star (see **Figure 3**). In each trial, the location of the stimulus varied randomly between four points around the center of the screen so that participants could not anticipate where the stimulus would occur in any given trial. The mask stimulus was a set of four rings, each surrounding a possible location of the stimulus (see **Figure 3C**).

Trials in which the stimulus was masked obeyed the sequence depicted in **Figure 4A**. In masked trials, the opacity of the target stimulus was lowered so that it could not be distinguished consciously (see **Figure 4A**). For each participant, prior to the main choice task, I employed a "staircasing" procedure to determine the specific opacity threshold at which the participant could no longer consciously perceive a masked stimulus. The procedure involves repeatedly increasing and decreasing the opacity of a masked stimulus until it is just below the opacity needed to be consciously visible to the participant (see **Figure 5)**.

In each trial of the staircase procedure, the participant was shown either a circle or a star, quickly followed by the mask stimulus, using the same stimulus and interstimulus time intervals as in masked trials of the main task. The participant was then asked whether the stimulus shown was a circle or a star, and they were also given a third option of "I don't know." The participants were instructed to select "I don't know" only if they had no intuition in either direction, such that it would be a pure 50-50 guess between the circle and star. To reduce the chance of false positives, participants were also asked to select the location where the stimulus appeared (top left, bottom left, top right, bottom right). If participants correctly selected the stimulus shape and location, then the opacity of that stimulus shape, but not of the alternate stimulus shape, was lowered by a factor of 0.9. Otherwise, the opacity was increased by a reciprocal factor of 1.11. A "reversal" occurred

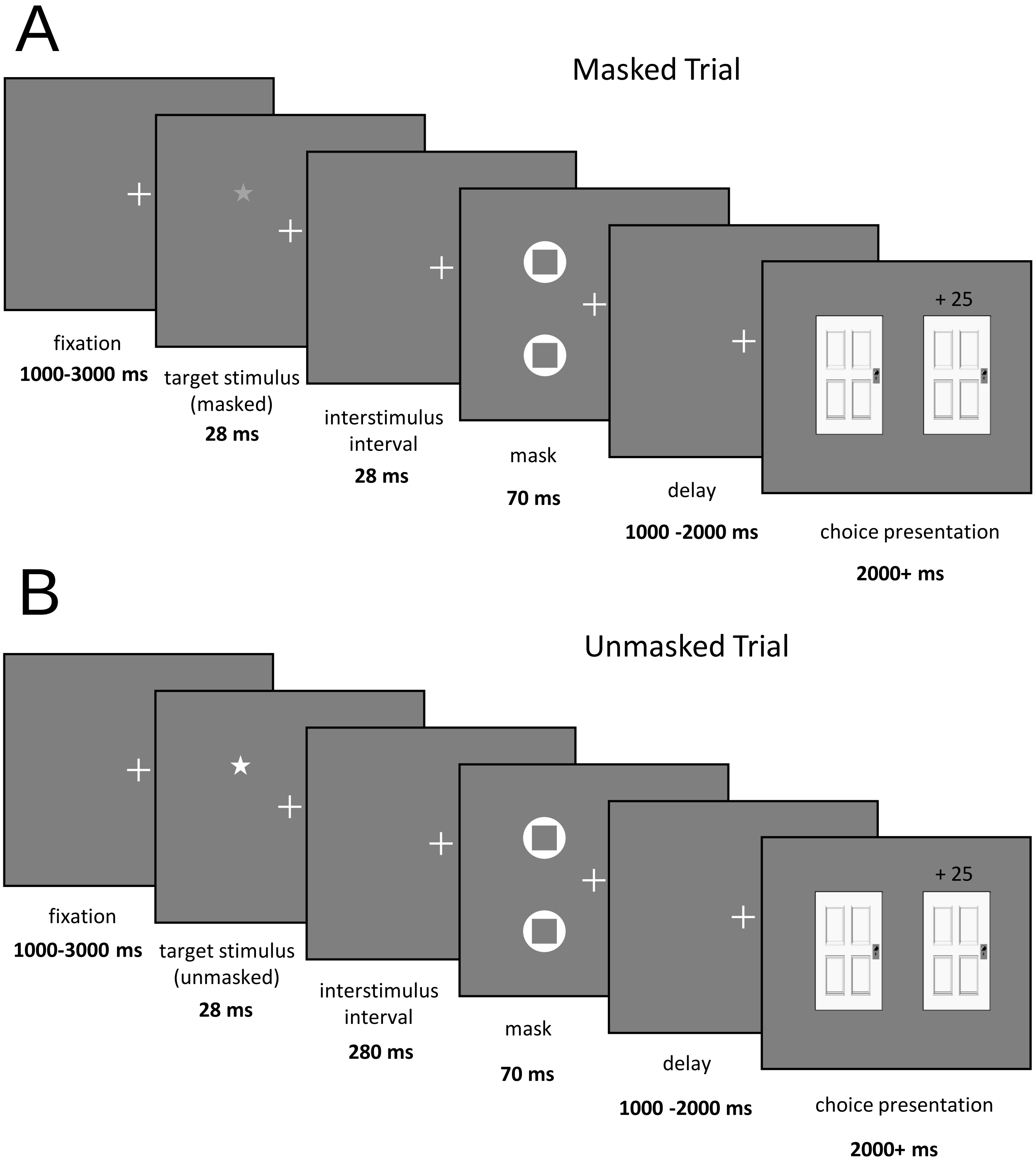


Figure 4 | Trial Timeline. Not to scale. **A.** Stimuli were masked by reducing both the target stimulus opacity and the interstimulus interval. The opacity level of the target stimulus in masked trials was determined by the threshold value obtained in the staircase procedure. **B.** In unmasked trials, the interstimulus interval was increased to 280 ms, and the target stimulus opacity was raised to 1. In both masked and unmasked trials, subjects were asked to choose between two doors at least two seconds after the choice was presented. Trials in the staircase procedure obeyed the same target stimulus, interstimulus, and mask stimulus time intervals as in the masked trials.

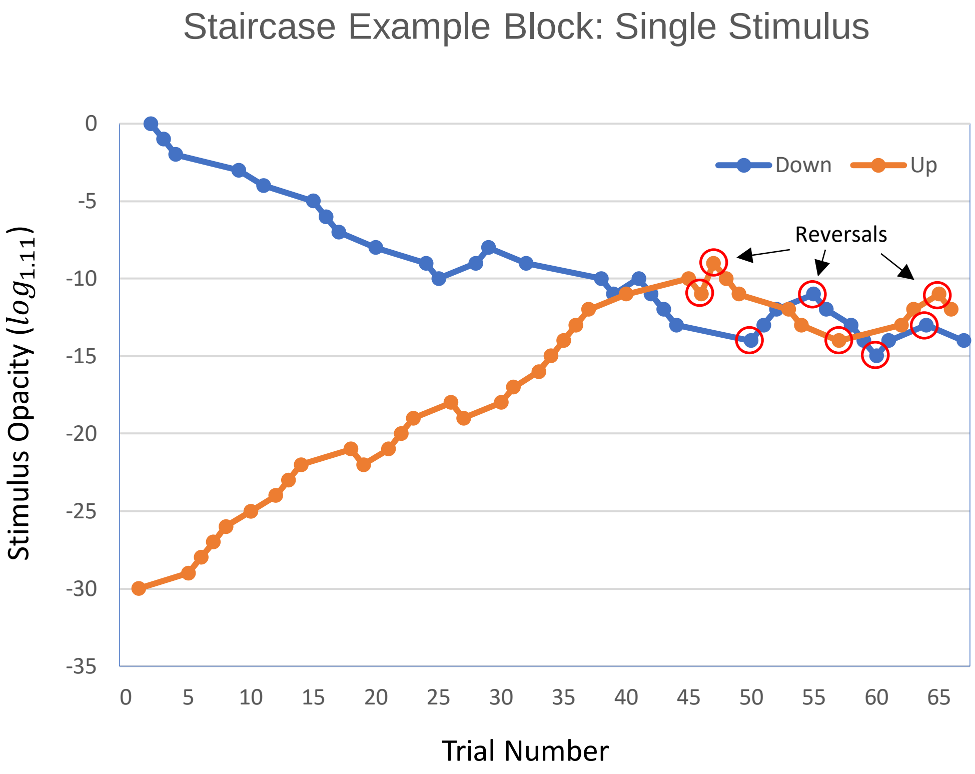


Figure 5 | Example Staircase Data. In the staircase procedure, stimulus opacity was multiplied by a factor of 0.9 whenever the subject correctly identified the stimulus and stimulus location, and opacity was multiplied by a reciprocal factor of 1.11 otherwise. The scale is logarithmic to correspond with the exponential step size. The staircase is performed in two directions: down (starting from maximum opacity of 1 = 1.11^0) and up (starting from a minimum opacity of 0.042 ≈ 1.11^{-30}). The four most recent reversals for each direction are circled in red. The opacity values at each of these reversals are averaged to determine the final threshold value. For simplicity, this example staircase data involves only one stimulus shape; however, in my actual staircase task, trials for both stimuli, in either direction, were randomly interspersed.

whenever the direction of opacity change switched from increasing to decreasing or from decreasing to increasing. Once eight reversals occurred for both the star and circle stimuli, the procedure was complete, and the threshold value was calculated by averaging the opacity values associated with the four most recent reversals. Separate threshold values were calculated for the circle and for the star.

Because the final threshold value may be influenced by the original opacity at which the stimulus is initially presented (Cornsweet, 1962), I ran this procedure, for both the circle and the star, in two different directions: starting both at the "top" of the staircase (opacity of 1.0) and at the "bottom" of the staircase (opacity of 0.042). Trials were randomized between the four "staircases" (circle starting at the top, circle starting at the bottom, star starting at the top, star starting at the bottom), and the threshold opacity for each stimulus shape was determined by averaging the two threshold values obtained from starting at maximum opacity and from starting at minimum opacity. If the variance of the four most recent reversal values of any of the four staircases was greater than 0.2, or if the difference between the threshold values when starting from the "top" and the "bottom" of the staircase was greater than 0.2, participants were asked to repeat the procedure.

Once threshold values were obtained, the main task of the experiment began. Prior to the task, participants were informed that they may earn a bonus of up to ten dollars depending on their choices during the task. Participants were also informed that their choices would determine the extent to which the bonus of the *next* participant in the study would be reduced, and similarly that their own bonus may be in part reduced due to the choices of the previous participant in the study.

In each trial, participants were shown a stimulus and then presented with a binary choice. The stimulus was either a circle or a star and was either masked or unmasked. If the stimulus was masked, its opacity was set to the threshold value determined by the staircasing procedure, and it was presented according to the same durations as in the staircase task (see **Figure 4A**). Meanwhile, if the stimulus was unmasked, its opacity was set to the maximum value and the interstimulus interval was increased from 28 ms to 280 ms to maximize its visibility (see **Figure 4B**).

After an additional 1000 ms, in both masked and unmasked trials, participants were presented with a choice between two doors pictured on the screen. One of the doors was visibly associated with a reward value that would be added to the participant's monetary bonus should that door be chosen (see **Figure 4**). The other door was not associated with any reward. At the start of the experiment, participants were informed that in any trial in which the stimulus presented was a *star*, selecting the reward-associated door would lead to a substantial risk of reducing the monetary bonus of the next study participant by an additional 2.5%. They were informed that this risk was present even in trials where they could not consciously perceive the star, so long as the star did appear. After the presentation of the doors, participants were made to wait two additional seconds before making their choice.

The forced choice task consisted of 250 trials in total, divided randomly between the four possible stimulus presentations (masked circle, masked star, unmasked circle, and unmasked star). To avoid decimals, reward values were presented in points, and every 100 points was equal to one cent. Rewards across trials ranged from 10-40 points (0.1-0.4¢), with the exception of half of the unmasked star trials, which were set to 2600-3000 points (26-30¢) in order to incentivize participants to choose the reward-associated door despite the risk of financial penalty to others.

III. Electrophysiological Recording and Analysis

Neural electrical activity was recorded using a BioSemi EEG amplifier and electrodes at sixty-four scalp sites. The ground and reference electrodes were located in the parieto-occipital area, one between electrodes PO3 and POz and the other between electrodes POz and PO4. Eye activity was recorded using electrooculogram electrodes placed around the eyes so that eye movement artifacts could be subsequently identified in the EEG data. Electrode impedance was kept below 20 kΩ, and EEG was sampled continuously at 2000 Hz (for more information on standard EEG procedures, see Light et al., 2010).

During analysis, the EEG data was resampled to 1000 Hz, re-referenced using an average reference across all electrodes to reduce signal-to-noise ratio (Verbaarschot et al., 2015), and bandpass filtered with a 0.1-50 Hz FIR filter. The epoch for ERP analysis was 3000 ms, starting 2000 ms prior to presentation of the choice and ending 1000 ms after the choice presentation. Highly noisy channels and epochs were removed from the analysis through visual inspection, and removed channels were interpolated using the data from nearby channels. The number of channels removed ranged from 0-2 (0.4 ± 0.7), and the number of epochs removed ranged from 8-65 (30 ± 23). The midline central electrode (Cz) was selected for statistical analysis based on prior studies (Chen et al., 2009; Polich, 2011).

For each subject, for each of the four conditions, all individual epochs were averaged into a single averaged ERP waveform. Amplitudes for these averaged waveforms were calculated using two different measures: mean voltage and area under curve (AUC). While the mean voltage method involves simply finding the average voltage value across a fixed interval determined in relation to stimulus onset, the AUC method is performed by adjusting the measurement window to the boundaries of the positive waveform, excluding any negative components, and calculating the total area between the curve and a selected baseline (see below). AUC analysis thus takes into account both the amplitude and the length of the waveform while avoiding the inclusion of neighboring negative ERPs (Luck, 2014). However, comparison of AUC values is also more dependent on the specific baseline selected than comparison of means is, as the choice of baseline affects not just the amplitude but also the length of the waveform being measured. For that reason, I employed both measures.

Before the amplitude was calculated using either mean voltage or AUC, a baseline correction was applied, wherein a specific time interval was selected and used to determine the new 0V baseline by subtracting the average voltage of that interval from the entire averaged epoch. The baseline for the mean voltage measure was determined by averaging the 200 ms interval prior to choice-presentation. Mean voltage for the P2 waveform was calculated in the time window of 150-250 ms after choice-presentation. Mean voltage for the P3 waveform was calculated in the time window of 300-400 ms after choice-presentation. Because of the potential effects of baseline choice in AUC analysis, AUC was measured with four different baselines, using the 200 ms, 400 ms, 600 ms, and 800 ms intervals prior to choice-presentation. The measurement window for AUC was 100-300 ms post choice-presentation for P2 and 250-600 ms post choice-presentation for P3. The curve for AUC was determined by identifying the largest continuous interval that contains the point of maximum voltage within the measurement window and which includes no more than five consecutive milliseconds below or equal to 0 mV. AUC was calculated by taking a Riemann sum of the curve in that interval (for an explanation of Riemann sums, see Oberbroeckling, 2021). If no such interval existed, meaning that there were no positive points in the measurement window, then the AUC was set to zero.

For the behavioral data analysis, each subject's rate of selecting the reward-associated door was calculated for each of the five conditions (the unmasked star trial was split into two conditions for the behavioral analysis: high-reward and low-reward). A one-way repeated measures ANOVA was performed to test the effect of trial condition on reward-selection rate, with each participant providing a reward-selection rate value for each condition. For the analysis of the electrophysiological data, a two-way repeated measures ANOVA was performed separately for P2 and P3 amplitude. The factors were stimulus shape (circle and star) and stimulus presentation (masked and unmasked). Each participant provided one amplitude value for each of the four conditions, which was calculated using either the mean voltage measure or the AUC measure. After it was observed that participants employed two divergent strategies during the experiment, strategy type was added as a between-subject factor in both the behavioral and ERP analyses.

Results:

I. Behavioral Results

First, I looked at the behavioral data to confirm whether the information about risk had a detectable impact on the choices of the participants. Across all subjects, the probability of selecting the reward-associated door was highest (99%) when the stimulus was an unmasked circle and lowest (57%) when the stimulus was an unmasked star with low reward value (see **Figure 6A**). The repeated measures one-way ANOVA, with Geisser-Greenhouse correction, showed a significant effect of trial condition (F[1.559,12.47] = 6.163, p = .018), though selected post hoc comparisons between conditions were not significant when adjusted with the Šidák multiple-comparisons correction (see **Table S1**).

The subjects' behavior during the experiment can be divided into two groups that employed divergent strategies (see **Figure 6B**). Five out of nine participants opted to select the reward-associated door in nearly all trials, including a majority of trials where the star stimulus was unmasked and the reward value was low. Four out of the nine participants opted for a more balanced strategy, regularly avoiding the reward-associated door when the star stimulus was presented and the reward value was low, but selecting the reward-associated door when the reward was high or the stimulus was a circle. This behavior is consistent with a strategy that aims to minimize risk of penalty for the future participant unless it is greatly outweighed by personal gain. I chose to categorize the subjects as employing a "reward-focused" strategy if they selected the reward-associated door in more than fifty percent of low-reward unmasked star trials, and otherwise I categorized their strategy as "balanced." A repeated measures two-way ANOVA revealed a significant effect of condition (F[4,28] = 23.774, p < .001), strategy (F[1,7] = 68.085, p < .001), and the interaction of condition × strategy (F[4,28] = 18.458, p < .001). Furthermore,

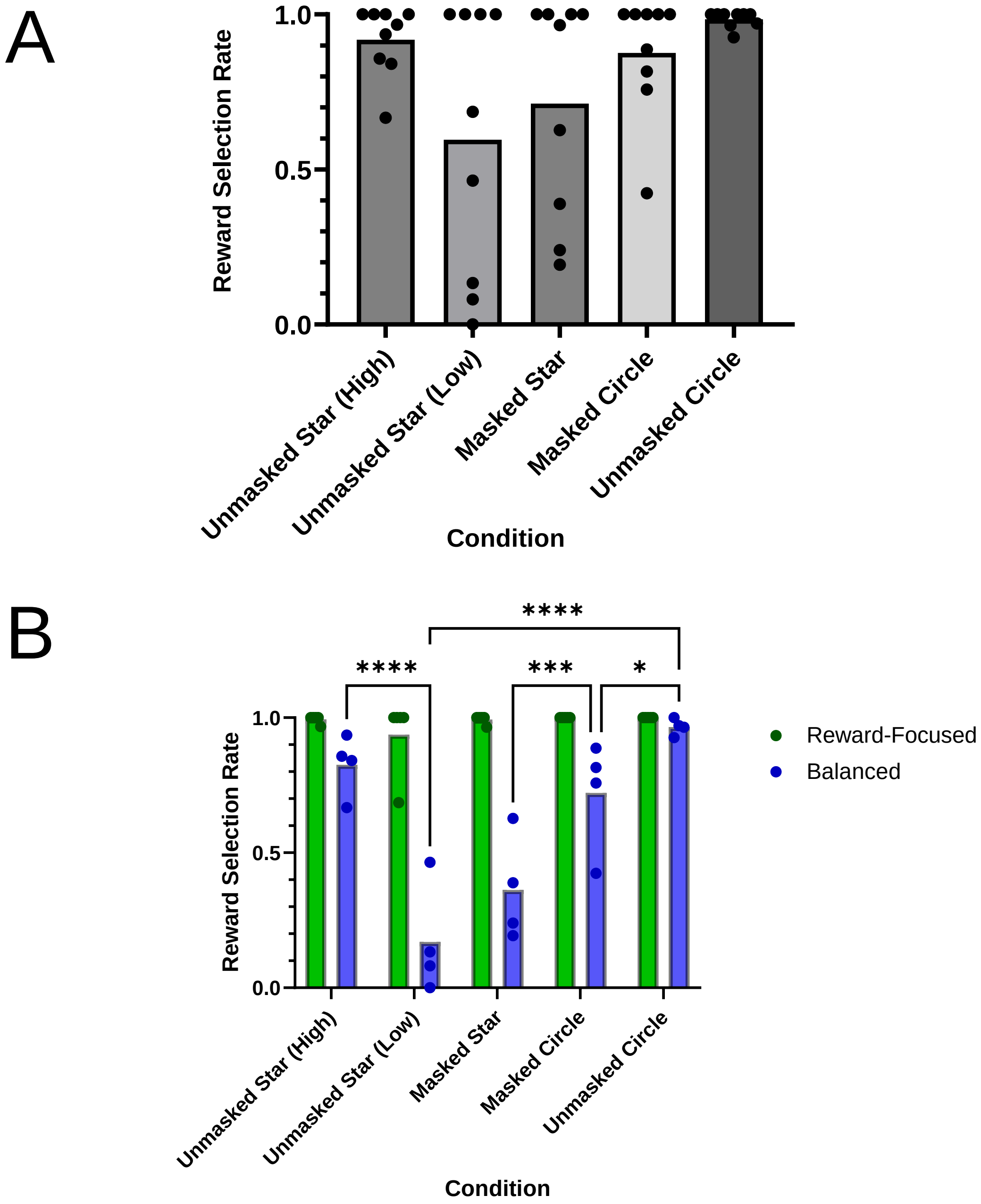


Figure 6 | Behavioral Results. A. Reward selection rates across the five conditions. The unmasked star condition is split into high and low reward conditions. **B.** I sorted the subjects into two strategy profiles, based on if they selected the reward in more than half or less than half of low-reward unmasked star trials. A two-way ANOVA revealed a significant effect of condition (p < .001), strategy (p <.001), and the interaction of condition × strategy (p < .001).

post hoc Bonferroni-corrected comparisons within strategy-groups revealed that participants were significantly more likely to select the reward option in high-reward unmasked star trials than in low-reward unmasked star trials (t = 8.586, p < .001), more likely in unmasked circle trials than low-reward unmasked star trials (t = 10.42, p < .001), more likely in masked circle trials than masked star trials (t = 4.702, p < .001), and more likely in unmasked circle than unmasked star trials (t = 3.198, p = .034). Comparisons were not significant within the reward-focused group (see **Table S2**).

II. *Electrophysiological Scalp Data*

The aggregated electrophysiological data at electrode Cz exhibited a small positive peak around 200 ms post choice presentation and a larger positive peak around 300 ms post choice presentation (see **Figure 7**). These peaks were more visually pronounced in subjects that employed the balanced strategy as opposed to the reward-focused strategy.

A. *P2*

The analysis of mean voltage for P2 showed no significant results for stimulus shape or stimulus presentation (see **Table S3**) but did reveal a significant interaction of stimulus presentation × stimulus shape (F[1,8] = 6.206, p = .037, see **Figure 8A**). The P2 AUC analysis results were highly dependent on the baseline employed (see **Table S3**). The ANOVA produced no significant results with the 200 ms baseline and a significant interaction of stimulus presentation × stimulus shape (F[1,8] = 6.390, p = .035) with the 400 ms baseline. There was a significant effect of stimulus shape (F[1,8] = 6.171, p = .038) and presentation × shape (F[1,8] = 6.559, p = .034), as well as an effect of stimulus presentation trending towards significance (F[1,8] = 5.203, p = .052) with the 600 ms baseline (see **Figure 8B**), and a significant effect of shape

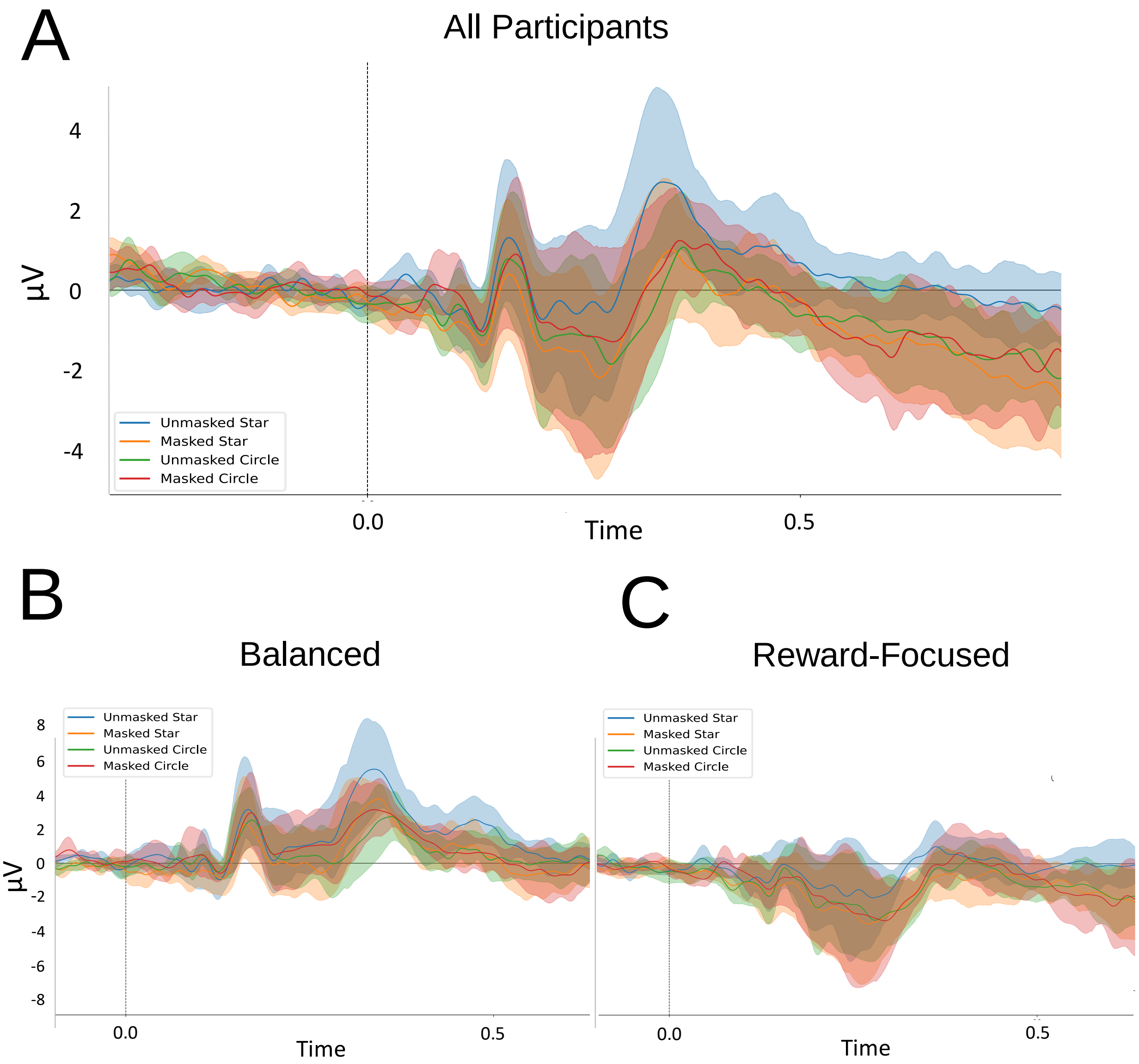


Figure 7 | Grand Average Electrophysiological Data. A. I aggregated subjects' electrophysiological data from the central midline (Cz) electrode, time-locked to the presentation of the choice, with 95% confidence intervals. Across all conditions, there were visibly distinguishable positive peaks 150-200 ms and 300-400 ms after choice onset. I also separated the averaged data into balanced (**B**) and reward-focused (**C**) strategy groups. The P2 and P3 waveforms were visibly more pronounced in the balanced strategy group data.

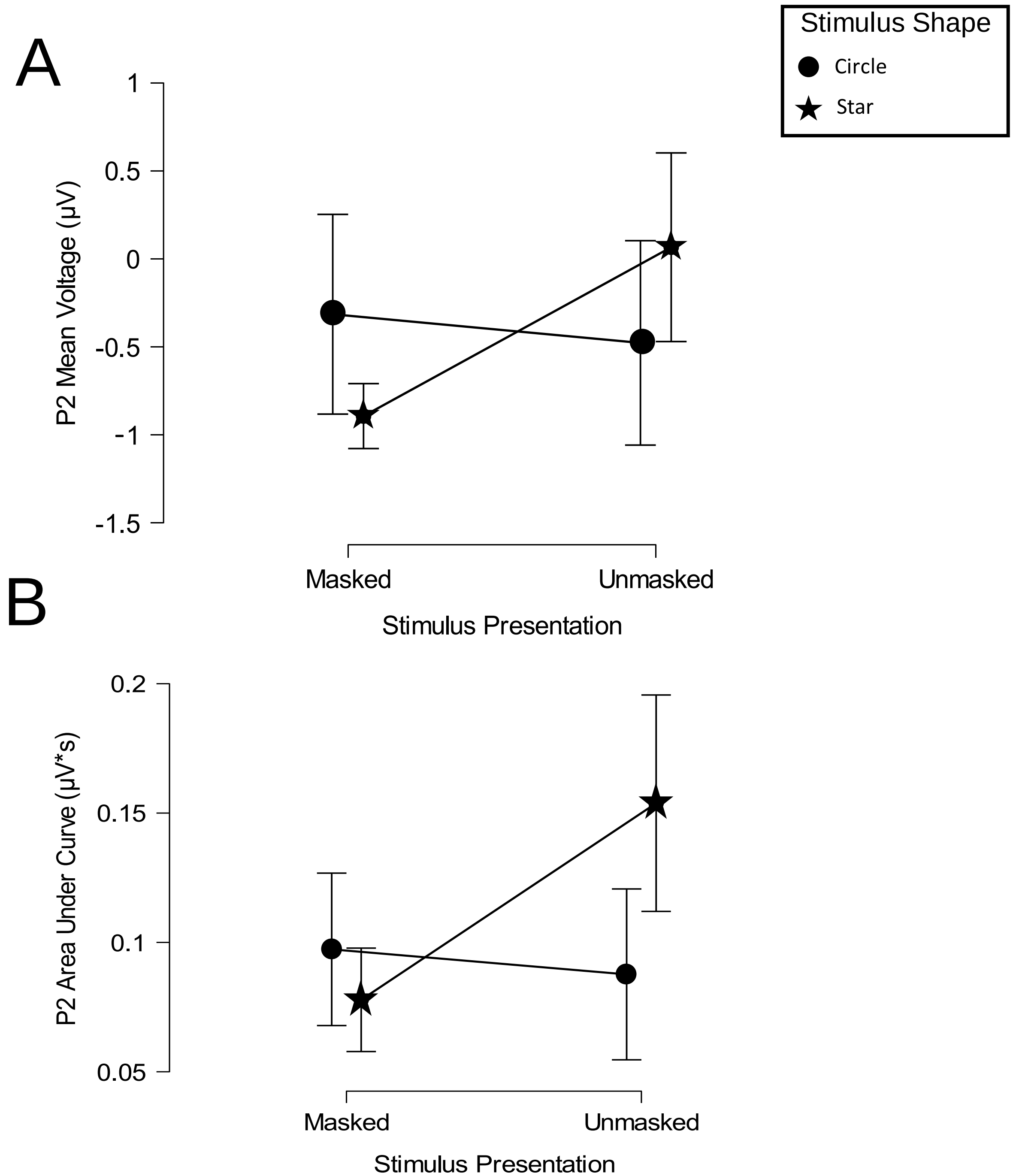


Figure 8 | P2 Amplitude with 95% Confidence Intervals. A. Mean voltage analysis of P2 amplitude yielded a significant interaction of stimulus presentation × stimulus shape (p = .037). **B.** AUC analysis of P2 with a baseline interval of 600 ms prior to choice onset revealed a significant effect of stimulus shape (p = .038) and presentation × shape (p = .034).

(F[1,8] = 5.947, p = .041), presentation (F[1,8] = 6.438, p = .035), and shape × presentation (F[1,8] = 6.523, p = .034) with the 800 ms baseline.

B. P3

The analysis of mean voltage for P3 revealed a significant effect of stimulus shape (F[1,8] = 26.080, p < .001) and shape × presentation (F[1,8] = 5.927, p = .041; see **Figure 9A**). AUC analysis was also dependent on baseline (see **Table S4**). There were no significant results with a baseline of 200 ms or 800 ms, although the effect of shape (200: F[1,8] = 5.174, p = 0.053; 800: F[1,8] = 4.859, p = 0.059) and shape × presentation (200: F[1,8] = 4.583, p = 0.065; 800: F[1,8] = 4.827, p = 0.059) were trending towards significant. With a baseline of 400 ms or 600 ms, there was a significant effect of shape (400: F[1,8] = 5.805, p = 0.043; 600: F[1,8] = 5.575, p = 0.046) and shape × presentation (400: F[1,8] = 5.455, p = 0.048; 600: F[1,8] = 5.604, p = 0.045; see **Figure 9**).

C. Analysis Including Participant Strategy

Because of the separation of the behavioral results into two distinct patterns based on participant strategy, and because of the corresponding difference in grand average waveforms (see **Figure 7**), I chose to add participant strategy as a between-subject factor of the ANOVA for P2 and P3 amplitude (see **Tables S5-S6**). With participant strategy included, the analysis of mean voltage for P2 had no significant results, although the interaction between stimulus shape and presentation was trending towards significance (F[1,7] = 5.358, p = .054). The interaction of shape and presentation was significant for all AUC baselines (200: F[1,7] = 5.598], p = .048; 400: F[1,7] = 13.410, p = .008; 600: F[1,7] = 15.688, p = .005; 800: F[1,7] = 15.256, p = .006), and the three-way interaction of shape × presentation × strategy was significant for the 400 ms, 600 ms, and 800 ms baselines (400: F[1,7] = 7.322, p = .030; 600: F[1,7] = 9.214, p = .019;

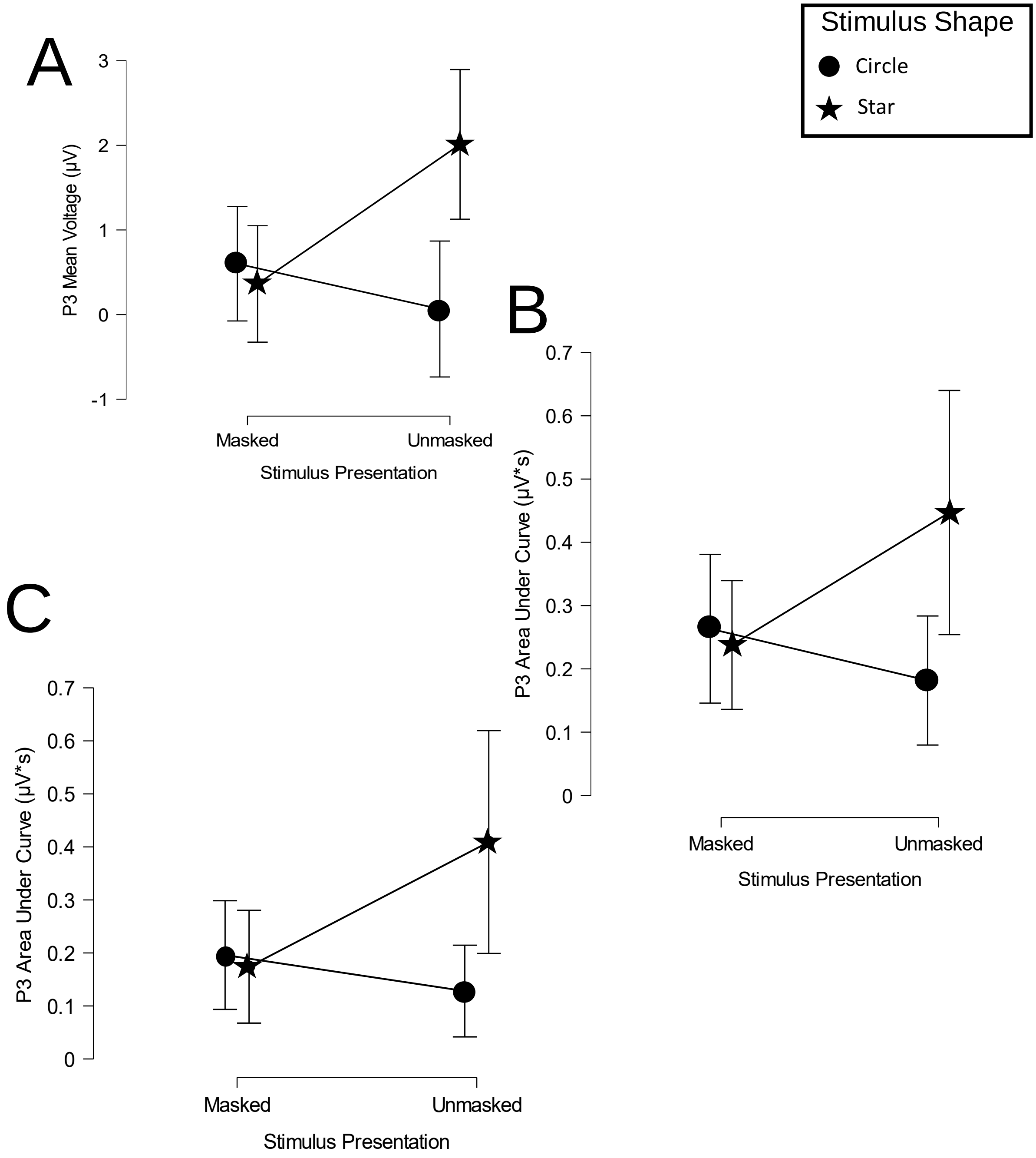


Figure 9 | P3 Amplitude with 95% Confidence Intervals. A. Mean voltage analysis of P3 amplitude showed a significant effect of stimulus shape (p < .001) and a significant interaction between stimulus shape × stimulus presentation (p = .041). **B.** AUC analysis with a baseline interval of 200 ms prior to stimulus choice yielded no significant results, although stimulus shape (p = .053) and stimulus shape × stimulus presentation (p = .065) showed a trend towards significance. **C.** AUC analysis with a baseline interval of 600 ms prior to stimulus choice revealed a significant effect of stimulus shape (p = .046) and stimulus shape × stimulus presentation (p = .045).

800: F[1,7] = 8.870, p = .021; **see Figure S1**). With a baseline of 400 ms, there was also a significant effect of participant strategy (F[1,7] = 6.125, p = .043). With the 600 ms baseline, there were additional significant effects of stimulus shape (F[1,7] = 8.333, = .023), participant strategy (F[1,7] = 6.394, p = .039), and shape × strategy (F[1,7] = 11.560, p = .011). Lastly, there were also significant effects of shape (F[1,7] = 8.899, p = .020) and presentation (F[1,7] = 6.831, p = .020) for the 800 ms baseline, and the effect of participant strategy was trending towards significance (F[1, 7] = 5.095, p = .059).

The mean voltage analysis for P3 showed a significant effect of stimulus shape (F[1,7] = 48.454, p < .001), participant strategy (F[1,7] = 16.318, p = .005), and the interaction between shape and strategy (F[1,7] = 6.821, p = .035; see **Figure 10A**), as well as a trending-significant interaction effect of shape × presentation (F[1,7] = 5.297, p = .055). AUC analysis with participant strategy included produced significant results across all baselines for shape (200: F[1,7] = 7.890, p = .026; 400: F[1,7] = 12.370, p = .010; 600: F[1,7] = 15.617, p = .006; 800: F[1,7] = 15.489, p = .006) and strategy (200: F[1,7] = 15.072, p = .006; 400: F[1,7] = 18.532, p = .004; 600: F[1,7] = 22.163, p = .002; 800: F[1,7] = 22.007, p = .002). The interaction of shape × presentation was trending significant for the 200 ms baseline (F[1,7] = 5.297, p = .057) and was significant for the remaining baselines (400: F[1,7] = 7.421, p = .030; 600: F[1,7] = 9.323, p = 0.018; 800: F[1,7] = 7.908, p = .026). Finally, there was a significant interaction between shape × strategy for the 400 ms, 600 ms, and 800 ms baselines (400: F[1,7] = 7.418, p = .030; 600: F[1,7] = 11.560, p = .011; 800: F[1,7] = 13.706, p = .008), as well as a three-way interaction effect of shape × presentation × strategy that was trending significant for the 600 ms and 800 ms baselines (600: F[1,7] = 4.479, p = .072; 800: F[1,7] = 4.207, p = .079; see **Figure 10B**).

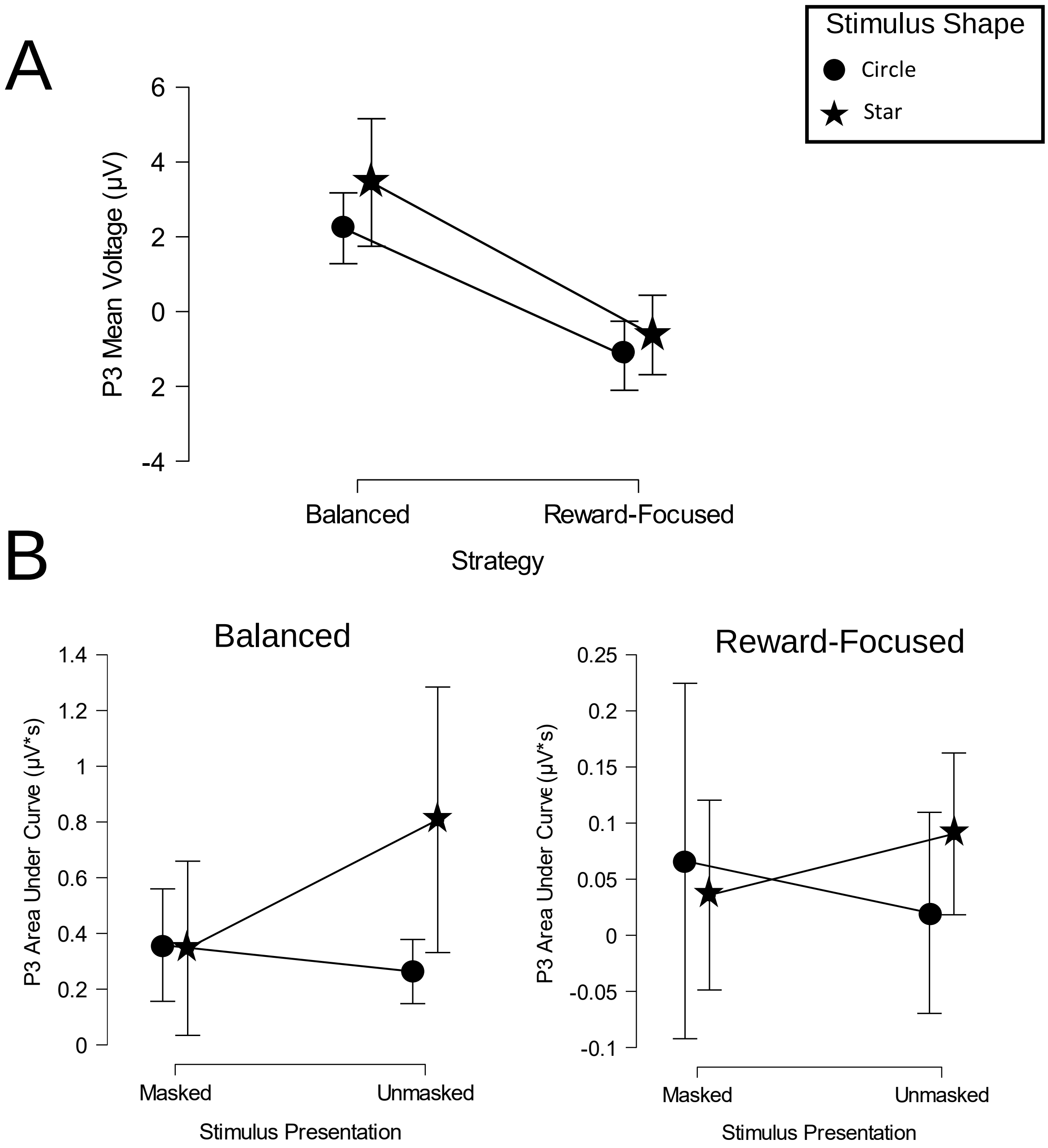


Figure 10 | Strategy Group Differences in P3 Amplitude with 95% Confidence Intervals. I included participant strategy as a between-subject effect in my analysis of P3 amplitude. **A.** Mean voltage analysis revealed a significant effect of stimulus shape (p < .001), participant strategy (p = .005), and the interaction between shape and strategy (p = .035), as well as an interaction effect of shape × presentation trending towards significance (p = .055). **B.** AUC analysis with a baseline interval of 600 ms prior to choice onset showed a significant effect of stimulus shape (p = .006), participant strategy (p = .002), stimulus shape × stimulus presentation (p = .018), and stimulus shape × participant strategy (p = .011). The three-way interaction effect of shape × presentation × strategy was trending significant (p = .072).

D. Reward Size

Because the reward value was large only in unmasked star trials, I conducted an analysis to rule out the possibility that the observed differences in P3 amplitude were caused only by differences in reward size. I removed trials in which the reward value was large (greater than 40 points), and with the remaining data I conducted a repeated measures ANOVA for mean voltage and AUC with a baseline of 200 ms (see **Figure S2** and **Table S7**). Both analyses showed a significant effect of shape (MV: F[1,7] = 12.249, p = .010; AUC: F[1,7] = 6.629, p = .037) and strategy (MV: F[1,7] = 16.516, p = .005; AUC: F[1,7] = 13.106, p = .009) similar to the analyses that included trials with large reward values.

Empirical Discussion:

My experiment provides preliminary evidence that conscious and unconscious processing of risk—specifically risk of harm to others—can be distinguished neurally, and that this difference in neural activity may reflect a difference in moral cognition. First, I successfully implemented an experimental paradigm in which information about the presence or absence of risk was presented both consciously and unconsciously. Participants opted for the self-benefitting yet risky option less frequently in trials in which the star stimulus was presented, suggesting that they were taking risk of harm to future participants into consideration in their decision-making. Furthermore, for individuals who employed a balanced strategy, there was a significant difference in reward-selection rates between circle and star trials when the stimuli were masked, although this difference was less pronounced than in unmasked trials (see **Figure 6B**). These results are consistent with a pattern in which the masked stimuli influence behavior without entering direct consciousness, as would be expected by a successful metacontrast masking paradigm.

The participants' factoring of risk of harm into their decision-making is mirrored by the relative amplitude of the P3 waveform at the central midline electrode. Using the mean voltage measure, I found that P3 amplitude was significantly greater in risky trials than in non-risky trials, and significantly greater in subjects who incorporated risk information into their decision-making than in subjects who instead opted uniformly for the reward. The AUC analysis matched these results when participant strategy was included as a between-subject factor, although it should be noted that when participant strategy was not included in the analysis, the effect of stimulus shape was only significant for two out of the four baselines I used, and trending towards significance for the other two baselines. Together, these exploratory analyses provide substantive, but not definitive, evidence that P3 amplitude indexes the incorporation of information about harmful risk into the participant’s decision-making.

Importantly, the P3 amplitude was selectively larger in risky than in non-risky trials when the stimuli were consciously presented, while it showed no such difference when the stimuli were masked (see **Figure 9**). This interaction between stimulus shape and stimulus presentation was significant for the majority of analyses conducted (mean voltage without participant strategy; AUC without participant strategy for baselines 400 and 600; AUC with participant strategy for baselines 400, 600, and 800) and trending towards significance in the rest of the analyses. The fact that such a difference was observed only in the unmasked condition suggests that the difference in P3 amplitude reflects a difference in risk level only when the information is presented consciously. Moreover, the lack of P3 amplitude difference in the masked condition despite the balanced group’s significant difference in behavior between risky and non-risky trials with masked stimuli (see **Figure 6B**) gives reason to believe that unconsciously presented risk information can affect behavior even if it does not lead to greater P3 amplitude.

One confounding factor that could complicate interpretation of the data is the possible effect of reward size. Specifically, in half of the unmasked star trials, I increased the reward size by two orders of magnitude in order to motivate subjects to choose the reward-associated door. As a result, it is possible that P3 amplitude, which has been found to be sensitive to reward value (Goldstein et al., 2006), was actually tracking the trial’s reward size rather than any moral features. However, the effects of shape and strategy on P3 amplitude remained significant even when only trials with low reward values were included in the analysis (see **Figure S2**), indicating that reward size cannot be the sole explanation for the differences in P3 amplitude.

It is important to emphasize that, because of the mix of highly overlapping yet slightly different analyses I conducted, these results should be taken only as preliminary. In my analyses of P3, the same tests were run on the same data, with either the measurement method being changed, the baseline correction being shifted, or a between-subject factor being considered. Accordingly, one should expect the results of these analyses to be highly similar but also acknowledge that there is an increased risk of a false positive result due to unanticipated effects of measurement or baseline choice. For that reason, these results cannot be taken as conclusive, but rather as providing motivation for further study.

The analysis for the P2 ERP amplitude was more mixed, as it was inconclusive regarding the effect of stimulus shape on P2 amplitude, and results were highly dependent on baseline. However, the P2 amplitude data is in part consistent with the difference in P3 responses between consciously presented and unconsciously presented stimuli. Specifically, there was a significant interaction between stimulus shape and stimulus presentation in the initial mean voltage analysis and all but one of the AUC analyses such that the P2 amplitude was higher in risky trials only when the risk information was presented consciously (see **Figure 8**). This result matches the pattern observed in the P3 data.

In short, there is promising yet non-confirmatory evidence that P3 amplitude tracks conscious representation of risk information. Given the previously established association of P3 presence with the processing of task-relevant stimuli and the evaluation of moral dilemmas, I propose that the increased P3 amplitude observed in unmasked star trials, compared to both unmasked circle trials and masked star trials, reflects an increased allocation of cognitive resources for the appraisal of stimuli indicating risk of harm to others and for the resolution of the moral dilemmas generated by these risk stimuli. As there is no moral dilemma when the circle stimulus is presented, P3 amplitude was comparatively lower in these trials, and as the moral features of the experiment did not influence the behavior of subjects who employed the reward-focused strategy, P3 amplitude was comparatively lower for these subjects. Finally, P3 amplitude was lower when risk information was presented consciously than when it was presented unconsciously, suggesting that subjects did not allocate the same level of cognitive resources to resolve the moral dilemma, even if the presence of the masked star stimulus suppressed their behavioral tendency to select the reward-associated door.

The neurobiological basis for the difference in P3 amplitudes may lie in the selective activation of temporoparietal, frontal, and subcortical regions involved in moral cognition and general decision-making. The P3 ERP is hypothesized to be generated from activity in temporoparietal areas, particularly those surrounding the temporal-parietal junction, as well as frontal areas including the lateral prefrontal cortex (Nieuwenhuis et al., 2005; Polich, 2007; Halgren et al., 1995). Research has shown that the temporal-parietal junction plays an important role in social cognition and moral decision-making, particularly when moral decisions involve one's self (Garrigan et al. 2016). The lateral prefrontal cortex is associated more broadly with executive control, decision-making, attentional allocation, and stimulus evaluation (Petrides, 2005; Yoder & Decety, 2018). Furthermore, some researchers have proposed that the P3 wave reflects activity in the subcortical locus coeruleus-norepinephrine system, which helps to realize decisions in response to motivationally salient stimuli (Nieuwenhuis et al., 2005). Thus, it is possible that the presentation of a more morally challenging choice would elicit a stronger P3 response, both because of the activation of neural structures with specific functions in moral reasoning, like the temporal-parietal junction, and also because the act of making moral judgments recruits general decision-making resources and many non-moral cognitive processes, such as attention, working memory, and emotion recognition (Garrigan et al., 2016; Yoder & Decety, 2018).

In the past, P3 has also been linked to activity in medial temporal structures (Nieuwenhuis et al., 2005; Polich, 2007; Key, 2005), which are strongly implicated in moral cognition (Garrigan et al., 2016). However, subsequent studies have shown that the P3-like activity generated in these deeper areas would not be able to directly account for the changes in voltage observed at the scalp recordings, though they may still have an indirect impact through their connections with frontal areas (Nieuwenhuis et al., 2005; Polich, 2007; Knight, 1984). Therefore, differential activation in these regions could most likely only play a partial role, at most, in causing the observed differences in P3 amplitude. P2, meanwhile, has been theorized to arise from medial frontal activity (Chen et al., 2009; Potts, 2004), and may relate to initial stimulus evaluation and detection of conflict (Lindholm & Koriath, 1985; Chen et al., 2009; Van Veen & Carter, 2002). Accordingly, differences in P2 amplitude could indicate differences in the detection and processing of moral conflict.

Though they are only exploratory in nature, my findings offer an initial case for a neurally distinguishable difference between conscious and unconscious risk awareness that appears to result from substantive differences in the way these types of information are processed in moral decision-making contexts. Consequently, these results are consistent with the position, implied by the standard legal definition of recklessness, that there is a distinction between the acts of disregarding consciously versus unconsciously represented risk.

However, these results should be accepted only with qualification. While my experiment has a number of strengths, including a real-life moral dimension that factored into participants' decisions and a paradigm that allows for both conscious and unconscious presentation of risk information, it also suffers from clear limitations. As mentioned above, I conducted several minorly diverging analyses, and not all were significant, meaning that I cannot definitively conclude that the results I describe are caused by something other than chance. Furthermore, my sample size was small—only nine participants, compared to the average ERP study size of 21 participants (Clayson et al., 2019)—and so to preserve statistical power, I only tested the electrophysiological data from one electrode (Cz). Analysis of data from neighboring electrodes would allow for a more robust set of conclusions. Additionally, the staircase methodology I used was imperfect, as participants frequently had to repeat the procedure after generating disparate threshold values the first time around, and participants were not retested at the end of the experiment to determine if their perceptual threshold had shifted over the course of the main task. Finally, there is an issue of interpretation: Since P2 and P3 correspond to a wide variety of cognitive processes, differences in their amplitude might be caused by some other feature of the task that is not morally relevant. For instance, since the unmasked star condition was the only condition with a large variance of reward value, it is possible that unmasked star trials more strongly activated regions that process novel and task-relevant stimuli, generating a larger P3 as a consequence of their variable reward size and not their moral significance.

The results of this experiment thus provide substantial motivation for further study. One future direction of research is to replicate this experiment while avoiding the limitations described above. For instance, researchers could avoid having to run multiple different analyses by testing out the different measures and baselines on an independent dataset beforehand, selecting the best analysis, and then performing the preselected analysis on a new set of participants. Researchers could also recruit a greater number of participants, providing sufficient statistical power to examine the data from a larger set of electrodes. Moreover, future experimenters could also improve on the staircase procedure by piloting the method on an independent set of participants in order to determine the optimal staircase parameters (e.g. step size, starting points, number of reversals) for identifying the participants' perceptual thresholds, as well as by testing participants' perceptual thresholds at the conclusion of the experiment.

In addition to addressing the methodological limitations, future research could also help clarify the interpretation of the data. If possible, it would be beneficial to conduct a version of my experiment in which reward value is not a potential confounding variable. More broadly, there is a need for additional research to determine the extent to which P2 and/or P3 reflect aspects of moral decision-making and social cognition, such that differences in P2 and/or P3 amplitude can be used to infer differences in moral cognition. This research could take the form of imaging studies identifying the neuroanatomical and functional sources of P2 and P3 or EEG studies involving moral dilemmas, such as Chen et al. (2019).

Finally, even if we could conclude with certainty that conscious and unconscious representations of risk involve different neural activity, that would not be sufficient to conclude that a moral or legal distinction between the two processes is valid. The next step would be to determine whether the neurally distinct representations of risk information are morally relevant. For this end, I now turn to a philosophical discussion on the nature of the relationship between conscious awareness and moral responsibility in general, as well as its implications for the assessment of reckless action.

Philosophical Discussion:

I. Setup

In the following section, I will discuss the relationship between conscious awareness and moral responsibility with the specific goal of assessing the claim that the former is a necessary condition for the latter. First, I will define the key concepts around which the discussion is centered, and I will outline three views that take different stands on the requirement of consciousness: volitionalism, the control view, and expressivism. I will describe a theoretical argument in favor of volitionalism and against expressivism and the control view, and then I will describe some cases that present a putative challenge to volitionalism. Next, I will explain why these cases are, in reality, not an issue for volitionalism and that our intuitions about these cases can be accounted for by making a distinction between the concepts of responsibility and ownership. Finally, I will apply this conclusion specifically to the moral concept of recklessness.

Before we can establish what is required for moral responsibility, we need to have a clear understanding of what the term means. Although there may be differences in the precise way in which competing theories define moral responsibility, they all generally understand it as a label that impacts how we ought to behave towards the agent in question. Neil Levy, for instance, states that assigning someone moral responsibility for an action means recognizing that "the fact that they have performed the action, in the circumstances and manner in which they acted, is relevant to how they may permissibly be treated when it comes to the distribution of benefits and burdens" (Levy, 2014, 2), while Angela Smith frames assigning moral responsibility for an event as "a basis for moral appraisal of that person" (Smith, 2005, 266-7), while leaving open the question of what that appraisal should be in each case. Although Levy is concerned with the allocation of rewards, punishments, and obligations and Smith is more broadly interested in the attitudes we form towards the person, both approaches take moral responsibility as having a role in how we treat an agent in a moral context. One might object that Smith's definition deals only with our *evaluation* of the agent, not our behavior towards them, but it is difficult to see how something could meaningfully affect our moral evaluation of an agent while having no effect on our behavior towards them in moral circumstances. For that reason, for the purposes of this discussion, I will rely on a general definition of moral responsibility according to which a person is morally responsible for an act or omission if their act or omission should factor into how we treat them in a moral context. It is certainly reasonable to disagree with this exact definition, but I suspect that the rest of my argument would be largely unchanged if adapted to a plausible alternative definition of moral responsibility.

Next, it is important to clarify which understanding of "consciousness" is relevant to this discussion. What is *not* relevant to the debates about moral responsibility is the type of consciousness associated with "qualia" or phenomenal experience, or the issue of how our physical brains can generate the subjective feeling of "what it is like" to experience one thing or another (as discussed, for instance, in Nagel, 1974, and Chalmers, 1996). Instead, the relevant sense is that of conscious *awareness*—that is, consciousness as a functional state "with an informational content" (Levy, 2014, 29). For instance, we are not conscious of actions we perform while asleep, and we do not consciously (but may unconsciously) process conversations occurring in the background while we are working on an unrelated task. A person can be conscious or unconscious of certain facts or experiences regardless of whether these facts or experiences have accompanying ineffable qualia, even if it means the person is what some have termed a "philosophical zombie" who lacks phenomenal experience.

II. Overview of the Debate

With our concepts of moral responsibility and consciousness pinned down, we can turn the discussion towards the meat of the matter: the extent to which conscious awareness of some aspect of one's action is a necessary condition of moral responsibility. On one side of this debate are theorists that can collectively be named "volitionalists." Volitionalist theories all contend that some level of conscious awareness is necessary for an agent to be morally responsible for an action. These theories can differ in their details—for instance, some are "choice"-based, stating that a person can only be responsible for an action or omission if it resulted from a choice they consciously made. I will choose to focus on a version of volitionalism that I will call the "fact-awareness" view. According to this view, what is required for moral responsibility is "consciousness of some of the facts that give our actions their moral significance" (Levy, 2014, 1). Two questions immediately arise from this definition: What does it mean for a fact to give an action its moral significance, and what does it mean to be conscious of a fact? While there are many potential answers to the first question—probably enough to comprise an entirely separate project—I will opt for the understanding that, under the fact-awareness view, an agent must be aware of facts which serve as at least *pro tanto* reason to believe that their action is wrong. For instance, under many moral theories, the act of theft is at face value wrong, even if there are instances where the wrongness of theft is outweighed by other moral reasons that make a thief's actions not immoral. The action of, say, moving a wallet from one place to another, however, is not wrong at face value, and only becomes wrong in the sense that it instantiates an act of theft. Consequently, being conscious of the fact that you are moving a wallet from place to place would not count as being aware of the facts that give your actions moral significance, but being conscious of the fact that you are stealing would be. Importantly, this general understanding of the facts that give an action moral significance is not dependent on the specific moral theory one endorses.

Let's now consider the second question. For Levy, a key proponent of the fact-awareness theory, an agent is conscious of a fact if the fact is "personally available" (Levy, 2014, 33) to them—meaning it is easily and effortlessly retrievable while also playing a role in the agent's occurrent cognition or behavior. This definition helps rule out cases where the information in question may be dispositionally available to a person but may not immediately come to mind without some sort of cuing—such as a person with dementia or in simple cases of absent-mindedness or forgetfulness—while also ensuring that the agent does not need to be continuously and actively thinking about the information at any given moment in order to be considered consciously aware of it.

In opposition to volitionalist views are those that contend that conscious awareness is *not* a necessary condition for moral responsibility, and that we can therefore be considered morally responsible for actions (or omissions) of whose moral relevance we are not conscious. These theories can be split into two general camps, which differ in the extent to which they broaden the scope of moral responsibility compared to volitionalism. The first camp, which only somewhat broadens the concept or moral responsibility, is that of "control"-based views, which state that a person can be responsible for an action or omission so long as they have "guidance control" over it (Fischer & Tognazzini, 2009, 38; Fischer & Ravizza, 2000). John Fischer and Neal Tognazzini, two advocates of this view, argue that this control requirement can be interpreted[3] as saying that

[3] Fischer and Tognazzini suggest that an alternative interpretation of control is that the agent has the "freedom to do otherwise" (Fischer & Tognazzini, 2009, 9). However, since they focus on a different interpretation based on reasons-responsiveness, and claim that these interpretations are interchangeable, I will address the reasons-responsiveness framing only. I suspect my argument regarding the freedom-to-do-otherwise view would be nearly identical.

an agent is in control of an action morally speaking if said action “issues from the agent’s own, moderately reasons-responsive mechanism. Roughly, the mechanism in question must be one for which the agent has taken responsibility and also one that displays a specific combination of receptivity and reactivity to reasons” (Fischer & Tognazzini, 2009, 9-10). This view does not require a conscious choice, or awareness of certain facts, for the possibility of moral responsibility; it requires only that an action or omission result from the properly reasons-responsive mechanism.

In addition to supporters of control-based views, there are those who contend that what matters in assigning moral responsibility for an action is the extent to which it reflects or expresses a person’s judgment. I will group these theorists under the label of “expressivists.” Angela Smith, a key proponent of expressivism, puts forth a view she calls the “rational relations” view. Under this view, an agent is morally responsible for something if the action or omission “reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it” (Smith, 2008, 369). As long as an action results from the agent’s evaluative capacities and attitudes, it need not be consciously apprehended by the agent in order for moral responsibility to be in the picture.

Expressivism consequently takes a broader stance on responsibility than most volitionalist views. It incorporates less clear-minded acts and omissions, like instances of forgetting, insensitivity, and carelessness, by giving more moral weight to the contents of unreflective thought. As Smith explains, “[i]f we value something and judge it to be worth promoting, protecting, or honoring in some way, this should (rationally) have an influence on our unreflective patterns of thought and feeling. We commonly infer from these unreflective patterns, or from their absence, what a person really cares about and judges to be important” (Smith, 2005, 247). Thus, for expressivists, what a person cares about and deems important, even implicitly, is a crucial component of how they should be morally evaluated.

The advantage of volitionalism over expressivism, and also control-based views, is its appeal to a neuroscientific understanding of the nature of consciousness, which allows it to convincingly challenge its rival theories from a theoretical standpoint. There are a number of competing theories attempting to explain how and why conscious awareness arises in the brain, but they all share the position that the function of consciousness is to unite and synchronize information across the brain—what Levy calls the “integration consensus” (Levy, 2014). The mind-brain is a somewhat modular system, organized into functionally-discrete processing components. Evidence from cognitive neuropsychology and functional neuroimaging has shown, for instance, that the brain processes conceptual information about animals and plants separately from man-made objects (Mahon & Caramazza, 2009; Farah & Rabinowitz, 2003), and that there are at least two distinct pathways for visual information, depending on if it is used for object perception (e.g. recognizing an apple) or to guide action (e.g. picking up an apple) (De Renzi, 2000; Goodale et al., 1991). But while many neural processes are domain-specific, rational information-processing occurs in a global, domain-general state. Even though our formulation of animate and inanimate objects may be functionally separable in their initial stages, at some point these concepts combine at the level of the entire person; in other words, we are not condemned to thinking about only one category or another at a time. Thus, the function of consciousness is to make information widely available across these semi-modular, dissociable systems so that they can work in conjunction.

One example of a theory explaining the integrative function of consciousness is the “global workspace” theory. According to this theory, consciousness broadcasts information to various component systems of the brain by means of a broad neuronal network distributed across the cortex, particularly in the prefrontal, temporo-parietal, and cingulate associative cortices. This network generates “long-range thalamocortical loops” (Levy, 2014, 49) that facilitate the constant updating of information and make this information available to a wide range of neural subsystems, which in turn guide behavior. There is a variety of neuroimaging, neuropsychological, and behavioral evidence that supports the idea of a global neuronal workspace or something similarly integrative. For example, conscious awareness is associated with increased activity in the same regions that contain high volumes of the pyramidal cells thought to be responsible for the extended thalamocortical loops of the global workspace (Levy, 2014; Dehaene, 2001; Dehaene et al., 2001; Laureys et al., 2002; Laureys et al., 1999). Not only are the relevant cortical regions correlated with consciousness, but so are patterns of coherence and synchrony between these regions (Levy, 2014; Fries et al., 1997; Srinivasan et al., 1999; Melloni et al., 2007; Gaillard et al., 2009; Gregoriou et al., 2009). Moreover, behavioral experiments meant to elicit unconscious processing using priming and cognitive load paradigms have shown that unconscious handling of information is more associative, stereotypical, and inflexible than conscious processing, leaving it unable to access logical properties of or relationships between stimuli (De Neys, 2006; DeWall et al., 2008; Baumeister & Masicampo, 2010; Hasson & Glucksberg, 2006). The incoherence of unconscious thought is consistent with the idea that the unconscious mind does not properly integrate information (Levy, 2014).

The reason the integrative, “broadcasting” function of consciousness is so critical for proponents of the fact-awareness view is that it allows for flexibility and reasons-responsiveness in action and thought, as opposed to the automatic mental scripts executed by local subsystems of the brain. In our motor behavior, we rely heavily on “schema”—rigidly defined routines of movement that are highly insensitive to environmental input. For instance, a tennis player who has mastered the perfect serve has likely developed a motor script for the serve that she can activate without having to (and probably without being able to) think about each individual step of the motion; it is a natural automatic movement. Similarly, an experienced driver in the midst of a conversation may be driving on "autopilot," responding perfectly to the requirements of the road just by relying on a set of well-learned action scripts—at least until he is faced with some surprising stimulus that returns his focus to the road (Levy, 2014). Importantly, these action scripts can also take the form of mental rehearsals of speech utterances, functioning as scripts of reflexive *thought*. Just as a tennis player can trigger an action script that goes through all the motions of a serve in an automated process, so do we have certain automatic thought patterns that may be activated by an external cue and which cannot be broken down into component steps. These mental scripts, whether they result in external behavior or internal thought, are highly insensitive to environmental stimuli, as they are essentially the products of subsystems of the brain implementing simple input-output functions. The gift of consciousness, thus, is to integrate information across subsystems in the brain so that these local processes are sensitive to a wider range of cues, allowing mental scripts to be modified or interrupted based on representations and attitudes that are held globally in the brain. Furthermore, by being able to override and modulate these automatic scripts, we can engage in the flexible process of reasoning that is important for moral responsibility.

As Levy points out, with this functional role of consciousness made clear, the basic setup of expressivism runs into problems. Expressivism as a theory turns on what exactly it means for something to "express an agent's evaluative judgment." Certainly not every behavior that is a causal result of a person's mental or neural processes is one that reflects their attitudes or evaluative outlook (an obvious example to the contrary is a seizure), so there must be a clear definition by which we can carve out the set of mental/neural processes and features that can be considered constitutive of one's rational capacities, morally speaking, versus those that cannot. However, no definition that Smith provides can avoid having consciousness play a functional role in the process of evaluative judgment.

For example, Smith explains that our judgments "taken together, make up the basic evaluative framework through which we view the world" (Smith, 2005, 251). While the judgments themselves may not be "consciously held propositions" (Smith, 2005, 251), the fact that they must be *taken together* to coalesce into one framework implies a level of coordination between the subsystems where individual judgments may develop—a coordination which, according to neuroscientific theories of consciousness, is facilitated by conscious awareness. She also claims that a morally attributable mental state "can reasonably be taken to reflect an evaluative judgment on the part of the person, a judgment, moreover, which it is appropriate, in principle, to ask her to defend" (Smith, 2005, 252). There are two issues with this claim. First, it is not clear how a judgment can be considered "on the part of the person" if it is not a "person-level" attitude accepted by the brain as a whole, rather than on a local level. Second, as Levy argues, the fact that the person should in principle be expected to defend an attitude suggests that it must be held consciously or broadcast globally—how could you defend or justify an attitude you picked up implicitly, an attitude you may not even realize you hold? Smith's description of evaluative judgment thus includes notions of integration ("taken together"), globality ("on the part of the person"), and rational justification ("to ask her to defend") that require a coordinating process like consciousness.

This trouble with expressivism can also be seen in Smith's discussion of instances that should not count as expressions of one's evaluative judgments. Smith places an emphasis on the sincerity of attitudes. For instance, she rules out the moral attributability of thoughts implanted by an evil scientist because the thoughts do not truly *belong* to the person—they are "induced in a way that bypasses her rational capacities altogether" (Smith, 2005, 262). Setting aside the complaint lodged above that *all* unconscious thought must inherently bypass one's rational capacities, there is another issue with this attempted distinction. It is not clear how an expressivist theory would be able to single out such clearly out-of-place cases from other instances of unconscious thought. Many of our implicitly held attitudes are forced upon us and bypass our rational capacities in some way, like those that are induced through social pressure or subliminal marketing campaigns. If an expressivist theory opts to include these cases, it cannot simultaneously exclude cases like the evil scientist without relying on an arbitrary criterion, such as a distinction between beliefs implanted through physically invasive versus non-invasive means. Even if a reformed expressivist theorist wished to also exclude cases of social pressure or manipulation, there would be no way to draw such a line non-arbitrarily considering the many unconscious beliefs that are acquired somewhere in the gray zone between pure, unfettered observation and nefarious string-pulling. Without a requirement that an attitude be broadcast and apprehended globally, expressivism welcomes any judgment that happens to be picked up by a local subsystem, having no way to differentiate between those that sincerely belong to us and those that affect our behavior in a manner that evades our rational faculties. Only conscious awareness, which integrates local beliefs into a coherent perspective, can ensure these beliefs are subject to our rational faculties.

A similar argument can be made in response to the control view of moral responsibility. Supporters of this view assert that an agent is morally responsible for an action or omission if they exert "guidance control" over it, meaning that it results from a moderately reasons-responsive mechanism. While the concept of a moderately reasons-responsive mechanism remains hazy, it includes a number of key features that require the integration of information that only consciousness can provide. For one, Fischer and Mark Ravizza, another advocate of the control view, emphasize that the reasons-responsiveness must be *moderate*, meaning it is not enough for the agent to be receptive to a very small number of reasons. Instead, the mechanism must recognize a "range of reasons" (Fischer & Ravizza, 1998, 81) to act, including "an appropriate range of moral reasons" (Fischer & Ravizza, 1998, 82), and its recognition of reasons must be "regular," such that it forms an "understandable pattern" (Fischer & Ravizza, 1998, 71). Thus, moral responsibility requires one's psychological mechanism for action to be receptive towards a sufficiently broad range of reasons and to have these reasons "connect and relate to one another" (Fischer & Ravizza, 1998, 72) in an appropriately coherent pattern—something which the unconscious mind is not equipped to do. Only the global workspace, or a similar instantiation of conscious awareness, can take in disparate points of information and incorporate them into a broader structure that follows a discernible logic. In fact, Fischer and Ravizza seem to directly acknowledge that moderate reasons-responsiveness must involve attitudes that are broadcast on the level of the full person, not just a local subsystem, as they state that the pattern of reasons-responsiveness for any given agent's mechanism could be deciphered, in theory, by holding an "imaginary interview" (Fischer & Ravizza, 1998, 71) with the agent in which they are asked what would count as sufficient reason to act given various contexts and various sets of values or preferences. A mechanism that agrees with the output of such an interview would necessarily reflect the explicit, global standpoint of the person and not include judgments that are held only unconsciously. It follows that guidance control would not be possible without fact-awareness. Levy's account of volitionalism thus appears to dismantle both expressivism and the control view on theoretical grounds, leaving it as, in principle, the only viable theory of moral responsibility.

However, the job is not done for the volitionalists. The advantage of the expressivist view is its ability to deal with a set of cases that pose a serious challenge to volitionalism, which I will call expressivism-friendly cases. These hard cases consist of actions or omissions that seem not to have conscious input but which nevertheless provoke intuitions of moral responsibility. One notable example is the case of the birthday forgetter, as described by Smith:

> *"I forgot a close friend's birthday last year. A few days after the fact, I realized that this important date had come and gone without my so much as sending a card or giving her a call. I was mortified. What kind of a friend could forget such a thing? Within minutes I was on the phone to her, acknowledging my fault and offering my apologies. But what, exactly, was the nature of my fault in this case? After all, I did not consciously choose to forget this special day or deliberately decide to ignore it. I did not intend to hurt my friend's feelings or even foresee that my conduct would have this effect. I just forgot. It didn't occur to me. I failed to notice. And yet, despite the apparent involuntariness of this failure, there was no doubt in either of our minds that I was, indeed, responsible for it. Although my friend was quick to pardon my thoughtlessness and to dismiss it as trivial and unimportant, the act of pardoning itself is simply a way of renouncing certain critical responses which it is acknowledged would, in principle, be justified"* (Smith, 2005, 236).

In this story, Smith has done something to hurt her friend's feelings—or more accurately, she has failed to do something and hurt her friend's feelings as a consequence. But as she notes, she never intended to avoid wishing her friend a happy birthday, nor did the fact that it was her friend's birthday ever enter her conscious awareness until days later. Nonetheless, there is a strong urge to believe that she is at fault and should apologize.

This type of case is difficult for a volitionalist view to account for because the seemingly morally responsible agent is not conscious of the facts that give her omission moral significance at the time of said omission. But beyond its evasion of a satisfactory volitionalist explanation, the reason expressivism-friendly cases pose a particular challenge is that the most natural explanation for our intuitions of moral responsibility in such cases is an expressivist one. It seems like the actual reason Smith is at fault for forgetting her friend's birthday is because such an omission reflects a miscalibration of her evaluative judgment and priorities, in the form of an internal lack of regard for a close friend. Something that evidently *wasn't* important to her *should have been* important to her, and that is why she is at fault. Since this explanation relies on us interpreting the wrongful acts as an expression of the agent's underlying values, attitudes, and evaluative judgment, and does not imply any conscious awareness or exercise of volition, a successful volitionalist account of moral responsibility must do one of two things. It would need to either explain why the intuitive judgments of moral responsibility in these cases are wrong or provide a volitionalist justification for such judgments to replace the expressivist one.

The standard volitionalist answer to these cases aims to do the latter. Specifically, it hitches moral responsibility not on the acts themselves, which were unconscious, but on previous conscious actions that led to the acts. Under this "tracing" view, a person is morally on the hook for a non-conscious action specifically because said action is a reasonably foreseeable result of a prior action for which they were conscious in the morally relevant way. The classic illustrative example of this type of explanation is in the case of drunk driving. If somebody becomes severely intoxicated and then decides to enter their car and drive home, they will not be presently conscious of the extreme risk they are undertaking, nor the harm they would cause should they hit someone, but we nevertheless hold them accountable for these consequences because, by consciously choosing to drink so much (and putting themselves in a position where they would need to drive home afterwards), they voluntarily initiated a series of events that they knew, or should have known, could lead to horrible consequences. Despite their present mental incapacitation, we can *trace* their moral responsibility back to these earlier choices.

The logic of tracing helps volitionalist views deal with a number of challenging cases, and some theorists have used it to explain expressivism-friendly cases like the three I outlined above. John Martin Fischer and Neal Tognazzini, for instance, argue that the reason Smith is at fault for forgetting her friend's birthday is because she "failed to take the necessary steps that any friend would take to remember friends' birthdays" (Fischer & Tognazzini, 2009, 37), like putting it in her calendar or setting a reminder for herself. Thus, she should be apologetic not for forgetting, but for putting herself in a position where she might forget, which Fischer and Tognazzini argue she did have conscious control over. The tracing method thus attempts to reconcile the expressivism-friendly cases with volitionalism by anchoring them to previous actions or omissions of which the agent did have sufficient conscious awareness or control.

III. Discussion

Despite its appeal, I contend that tracing is an inadequate solution to the class of expressivism-friendly cases for multiple reasons. For one, it is difficult to believe that for every product of unreflective thought that we think merits moral responsibility, there is some past action or omission that meets the requirement of consciousness. Let us start with Smith's case of the birthday forgetter. If, at some previous moment, Smith had actively made the decision not to set a reminder for her friend's birthday—perhaps thinking to herself, *I know this might lead me to forget my friend's birthday, but I will avoid setting a reminder nonetheless*—then the case for tracing would be much more clear. But this sort of explicit decision is unlikely. More likely, the idea of setting a reminder had never even occurred to her, or it was something she had planned to do and forgot. In these cases, it seems like tracing can only push the problem of forgetting back, so that we must search for some previous conscious act or omission that explains why it never would have occurred to Smith to take the necessary steps to prevent herself from forgetting a friend's birthday.

At this point it is worth noting that the fact-awareness view is not the only view that attempts to explain expressivism-friendly cases in terms of tracing. In particular, proponents of the control view like Fischer and Tognazzini argue that their theory is preferable over a choice-based or fact-awareness theory specifically because it makes tracing easier: It is easier to find previous actions or omissions over which we had control in this broad sense than it is to find instances of conscious choice or fact-awareness. Of course, this argument is obsolete if we accept the theoretical case against the control view outlined above; but even if we allow that guidance control does not require fact-awareness, and even if we concede that Fischer and Tognazzini are generally right about the advantage of the control view in finding traceable events, even this view cannot explain all the expressivism-friendly cases.

The control view of tracing is specifically hard to formulate when the event in question is traced back to a previous omission rather than a previous action. In such cases, the past omission in question must result from a sufficiently reasons-responsive mechanism, which immediately raises questions about how the *lack* of an action can be the result of a psychological mechanism. Fischer and Ravizza define responsibility for omission in a manner that is "symmetric" (Fischer & Ravizza, 1998, 132) to the control view's version of moral action. They reformulate omission as "the agent's bringing about relatively finely-specified negative consequence-universals" (Fischer & Ravizza, 1998, 134)—or in more comprehensible terms, they hold the agent responsible not for the omission per se, but for causing, through a different set of actions, an outcome in which the results of the omitted action do not occur. For instance, if a person witnesses a child being kidnapped but does nothing to stop it, they are responsible specifically for *staying put in such a way that the child is not saved from kidnapping by them*. Moreover, in order for reasons responsiveness to hold for omission of action X, not only does the agent's act of staying-put-such-that-X's-consequences-do-not-occur (which I will call action Y) have to result from a mechanism that is receptive to reasons, including moral reasons, in a coherent pattern, but there must also be some “alternative scenario” (Fischer & Ravizza, 1998, 138) within this pattern in which the agent has sufficient reason to act otherwise than Y (i.e., by doing X), and consequently does so.

When it comes to tracing, there are two problems with this understanding of omissions. First, the action Y becomes undefined as soon as we consider cases of omission that are not fixed to a specific moment in time. Y is easily describable in the kidnapping example because the agent’s opportunity to do X is limited to a brief, discrete, interval. But what about in cases like the birthday forgetter, where the agent had many continuous opportunities to set a reminder for herself? Since X could have occurred at countless points throughout an extended time period—say, a year—is Y then the set of *all* the agent’s actions within that year, since they together constitute staying-put-such-that-a-birthday-reminder-is-not-set? This consequence of Fischer and Ravizza’s view is bizarre, as it implies that when we say an agent is responsible for not setting a reminder, we are really saying they are responsible for everything they did in the course of a year when they could have been setting a reminder.

The second, possibly more crucial, problem, is that there are many cases in which action X is so foreign to the agent that it simply could have never occurred to the reasons-responsive mechanism at all, even unconsciously, and so there is no alternative scenario in which the agent has sufficient reason to perform X. For instance, in the example of the birthday forgetter, the idea of setting a birthday reminder could have been totally alien, or at least unintuitive, to Smith, such that the potential course of action fell outside the scope of Smith’s reasons-responsive mechanism and it was not even a live option for her. One might be tempted to argue that there are still scenarios in which it is a live option for her—like those in which the benefit or importance of setting birthday reminders is made clear to her—but counting these kinds of scenarios as valid alternatives would make responsibility for omissions far too broad, putting us on the hook for omitting all sorts of things of which we are totally ignorant or which never factored into our conscious or unconscious reasoning whatsoever. Therefore, we must concede that even under the control view, we are not responsible for some omissions that are supposed to anchor tracing. What's more, this approach to tracing seems morally irrelevant: Not only would Smith's friend have no way to know whether Smith did not set a reminder because some (conscious or unconscious) reasoning process had passed over the idea or instead because it was never in Smith's sample space of actions to begin with, but it is also doubtful such a distinction would make a difference to the validity of the friend's grievance.

In short, whether one adopts the fact-awareness view or a control-based view of moral responsibility, tracing leaves behind gaps and mysteries. But even if one is not swayed by the problems brought up in Smith's birthday case, the opaqueness of tracing is an even greater issue in other types of expressivism-friendly cases, like instances of insensitivity and microaggression. Consider two examples of mine below:

1. *The Regretful Laugher*

Suppose Shawn is going out to lunch with a group of friends, and they are in the midst of a lively conversation while they wait for their food. During this conversation, Shawn's friend, Kelly, makes a joke at the expense of his other friend, Ned, targeting a big (and widely known) insecurity of his: his hairline. Without thinking, Shawn laughs at the joke before immediately stopping himself—but it is too late. He turns to Ned and sees he is deeply hurt, not just by Kelly but by Shawn, for laughing at his expense. Shawn feels awful. He is sure that he would have been stone-faced and disapproving of the joke had he fully processed what Kelly was saying, but he was so caught up in the flow of the conversation that his laugh was practically involuntary. Still, he immediately tries to find a way to make it right with Ned, and to let Kelly know it is not okay to make jokes about others' insecurities.

Unlike in the first case, the case of the regretful laugher features an *action*, rather than an omission, but once again, it seems that the agent has done something wrong and should be subject to some form of moral appraisal—even though the act he committed had no conscious input. When Shawn laughed at the joke, he was not consciously aware that it might hurt Ned's feelings, but he still feels guilty for doing so afterwards. This case is thus another challenge for volitionalism. It is worth noting that there is also a potential (non-tracing) control-based account of responsibility in this case, as it is possible that the mental mechanism by which Shawn evaluates humorous statements and laughs in response is sufficiently reasons-responsive to merit moral responsibility.

2. The Microaggression

Arthur is a white college professor who teaches a comparative literature course of about twenty students. In a lecture about halfway through the semester, he calls on one of his students, Lisa, but mistakenly addresses her as "Sarah," the name of another student in the class. Both Sarah and Lisa happen to be Asian. Arthur continues his lecture, unaware of his mistake, but Lisa and Sarah are both embarrassed by the mix-up.

In this case, Arthur does not intend to confuse his student's names, nor is he consciously aware that he is going to make such an error—in fact, he is not even aware of the mistake after the fact. No aspect of the gaffe ever enters his consciousness. And yet, there is still an urge to condemn his actions or hold him at fault. This sort of case also requires a volitionalist explanation.

By now, you may be able to anticipate how, in the broadest terms, a tracing explanation might approach these examples. In the case of the regretful laugher, a proponent of tracing might argue that the only reason Shawn was unintentionally inconsiderate of his friend's insecurity in the first place is because of some previous freely-committed actions or omissions that led him not to care enough about his friend's sensitivities, or which made him disposed to finding hurtful jokes worth laughing at. And in the case of the microaggression, Arthur's name mix-up, despite his obliviousness to it, may possibly be traced back to previous failures to confront his biases or previous instances where he allowed himself to be receptive to racist attitudes.

Nevertheless, these explanations become elusive once we step out of the abstract. In the case of the regretful laugher, it is hard to conceive of any specific actions that could have prevented Shawn from laughing at the joke, particularly any actions that he could reasonably have known would prevent him from doing so. At best, there is a vague set of behaviors that Shawn should have espoused more firmly that would have allowed him to cultivate the attitudes needed not to laugh at the joke—such as listening to the experiences of those dealing with aesthetic insecurities or thinking critically about the effects of hurtful jokes on others—but it is challenging to contend that he should have foreseen this incident (or ones like it) as an expected consequence of neglecting to do these things, or that he was aware of the facts giving moral significance to not engaging in these behaviors at the time he was failing to engage in them. Likewise, in the case of Arthur's microaggression, there is no specific preventative measure he could have been expected to take, as he was totally unaware that he commits these sorts of microaggressions to begin with. Instead, Arthur needs to have operated the course of his life in such a way that he does not implicitly view Asian people as interchangeable, or is attuned to how his bias can shape his day-to-day interactions with others, or at least has acquired the habits or values that allow him to be more mindful about how he addresses the Asian students in his class. Even more knottily, it is possible that the only way Arthur could have prevented a mistake of this sort was by growing up around more Asian people—something entirely out of his control. Once again, while it is preferable for Arthur to have developed in any of these manners, there was likely no moment or set of moments where Arthur stood at the figurative crossroads between microaggression and non-microaggression and was conscious of the facts giving his actions moral significance in this respect, or even in control in a reasons-responsive way. Note that this problem is not merely epistemic—it is not that we can't *know* the event or events, for which Arthur can be held morally responsible, which can serve as the basis for tracing back his microaggression. Rather, the event or events may not exist at all.

This argument is similar to one made by Manuel Vargas in "The Trouble with Tracing," in which he makes the case that the knowledge condition is not always met (i.e., the results of one's actions are not always reasonably foreseeable) in the original acts we trace back to. The point I am making, however, is even broader. Not only are there original acts where the knowledge condition is not met, but there are also those in which the agent has no *awareness* of the facts that give their action moral significance, or does not act deliberately, or has no control over the situation at all. In all such cases, there is, metaphysically speaking, nothing to draw the line of responsibility back to. Tracing can fail in a number of ways.

Even if we do buy that all expressivism-friendly cases are traceable to a morally attributable past action or omission, there is still a second problem. The driver of our moral intuitions about moral responsibility in these cases is still the act itself, and what it reflects about the person, not the past actions that led to the event in question. If you are upset that a friend forgot your birthday, you are likely hurt that they didn't value you enough to remember, not that they didn't set a reminder for themselves. Suppose, for instance, that Smith had set multiple reminders about her friend's birthday, but that it still slipped her mind to call her friend when the opportunity was right. In this case, it seems like the friend would still have reason to be upset with Smith (although perhaps less upset), and Smith would still have reason to apologize. Or conversely, suppose that Smith makes a new friend, and this friend mentions the date of their birthday in passing during one of their early conversations, and suppose that when that day comes, it occurs to Smith to wish this friend happy birthday, even though she had set no reminders and was not explicitly trying to remember the birthday since that conversation—it had simply occurred to her. In this case, it seems that Smith has earned legitimate praise for thinking of her friend, even though she took none of the necessary steps outlined by Fischer and Tognazzini. Or perhaps most strikingly, suppose that Smith had actually remembered her friend's birthday in the original example, although she had not set any reminders for it. It would be absurd to find her at all *blameworthy* for not setting any reminders in the face of the fact that she did indeed remember the birthday. These examples should draw out the intuition that the central issue in the forgetting case is whether Smith cares sufficiently about her friend, not whether her forgetting or remembering was facilitated by some previous omissions or actions.

This criticism applies even more strongly for the other two cases. Ned is upset at Shawn, not for historical choices Shawn made that prevented him from cultivating a proper sense of empathy or considerateness, but *for thinking those kinds of jokes are funny or okay*, and for laughing as a result. Similarly, Arthur's students likely resent him for not seeing them as individual people and consequently humiliating them, not for the numerous subtle moments throughout his life that may have brought him to that moment. In short, the moral concerns in cases like forgetting, insensitivity, and microaggression, are the priorities and attitudes reflected by the person's unconscious action, not the conscious actions or omissions to which the unconscious acts are supposedly traced.

Thus, we are left with a conundrum. On the one hand, the fact-awareness view deconstructs (or at the very least absorbs) expressivism on theoretical grounds—there is no expression of evaluative judgment without global broadcasting of information. On the other hand, there are expressivism-friendly cases that seem to 1) lack the qualities necessary for moral responsibility under the fact-awareness view, 2) warrant judgments of moral responsibility that are supported largely by expressivist intuitions, and 3) evade tracing. So how do we proceed? I propose that, rather than validating the intuitions of moral responsibility in these cases by attempting to trace them, we should show why they are false intuitions to begin with. The key to doing so is by distinguishing between the concepts of responsibility and ownership.[4]

[4] In their discussion of the control view, Fischer and Ravizza also discuss a concept they also call "ownership." The notion I discuss here is different from theirs.

To explain this distinction, let us consider the case of an unkind sleepwalker. Suppose Kevin is a college student who lives in an apartment with two other students. They share a kitchen but each person buys their own groceries and cooks their own meals. Now suppose one night, while he is fast asleep, Kevin gets out of bed, walks to the kitchen, takes a slice of cake from the refrigerator that his roommate, Carl, had been saving for the next day, and tosses it in the garbage. Carl, who is a late owl, happens to be in the living room at the time and witnesses the whole thing. Carl confronts Kevin about it in the morning, much to Kevin's surprise.

In my view, there are three features of a natural response to this parable, of which I will initially discuss just two. First, it should be obvious that Kevin is not morally responsible for throwing out Carl's slice of cake, since it was not an act he had control over or awareness of, nor a reflection of his evaluative judgment.[5] Still, I expect that the second feature of our reaction is the understanding that it makes sense for Kevin to feel apologetic or for Carl to be upset with him, even though morally speaking Kevin has done nothing wrong. These two points together illustrate the notion of *ownership*. Although they do not result from conscious deliberation or a reasons-responsive mechanism of his, in some sense, the actions *belong* to Kevin. They are a part of who he is, even if they do not characterize his whole person.

These two features of ownership are similarly present in the notion of "agent regret" as described in the literature on moral luck. Agent regret is the sentiment that arises when one is *causally*, but not morally, responsible for some harm. The classic example is of a truck driver who, having horrible luck, hits and kills a child through no fault of his own. As Bernard Williams points out, the driver has a special relationship to this event that separates him from just any spectator.

[5] A bold expressivist could argue that Kevin's actions while asleep actually do express his evaluative judgment in some way, and so he is morally responsible for them. However, I take it that most people would find it absurd to make inferences about an individual's character based on their behavior while sleepwalking.

Although he is not to blame whatsoever, he will likely feel a horrible sense of guilt and ruminate on what he could have done to prevent the accident: “Doubtless, and rightly, people will try, in comforting him, to move the driver from this state of feeling, move him indeed from where he is to something more like the place of a spectator; but it is important that this is seen as something that should need to be done, and indeed some doubt would be felt about a driver who too blandly or readily moved to that position” (Williams & Nagel, 1976, 124). In other words, although there is no ascription of moral responsibility, there is an expectation for the driver to feel distinctly bad, at least initially, about his causal role in the event, to the point where it would be considered out of line for him not to express deep remorse and dread even though, plainly speaking, he has done nothing wrong. The case of agent regret demonstrates that immediate natural judgments of blame and guilt may be divorced from actual instances of moral responsibility. This same dissociation applies to judgments of ownership.

Another loose analogy where this concept of ownership sans responsibility may apply is the case of family history. Suppose that, prior to your birth, your family was complicit in a horrible crime that ruined several people’s lives (but, for simplicity’s sake, suppose that neither you nor any of your family members benefited from this crime). If, as an adult, you were to find out about these crimes, it would be understandable to feel—perhaps even abnormal not to feel—a sense of guilt and a desire to make things right in some way. However, no one could reasonably blame you for this crime, or even accuse you of profiting off of it. Nevertheless, your connection to the crime is still an aspect of your history and identity. You take ownership over it in the sense that it is an instinctive, emotionally understandable reaction to connect the wrongdoing to you without actually subjecting you to moral appraisal.

Or, finally, consider what often happens when a corporation experiences a major scandal or causes social harm, perhaps as result of a snowballing error that slipped through the bureaucratic cracks. In many cases, the harm in question was not the fault of the CEO of the company—possibly the fault of no individual at all—but rather a cumulative consequence of many small mistakes in lower levels of the company. Nevertheless, the CEO commonly resigns in such cases, or at least experiences pressure to resign. Despite not being morally responsible for the misdoing, there is a sense in which they must take ownership over it as the head of the company. Ownership in the expressivism-friendly cases works in a structurally similar way: Even though we are not morally responsible for the wrongs that slip through the cracks of our mental machinery, we as conscious beings are arguably like CEOs of those various unconscious subcomponents that constitute our mind, and thus we ought to express ownership over the harms that arise from them.

What's important about this notion of ownership is that it is nothing more than a characterization of a natural psychological response that *resembles* intuitions of blame. It explains why we may be tempted to assign moral responsibility in the expressivism-friendly cases, but it itself is not a moral label. However, not all of these cases can fully escape blameworthiness. That is where the third feature of our natural reaction to the unkind sleepwalker case comes in: If such an incident were to happen again, Kevin would no longer be free of moral responsibility, and he would be more responsible each subsequent time it occurs. This intuition can be explained by a standard view of tracing. Once it has been brought to his attention that he behaves in this way while he sleeps—that his unkind sleepwalking is a part of who he is—we can start to blame him for not taking the necessary steps to prevent such behavior (like locking the door or seeing a doctor), especially if it becomes a pattern. What starts out as mere ownership turns into a story of blame due to tracing.

Now that the concept of ownership is fleshed out, we can begin to apply it to the expressivism-friendly cases. In all three cases, our expectation that the misbehaving individual should be apologetic, and that the aggrieved individual is valid in feeling wronged, can be explained as intuitions of ownership. Just like the unkind sleepwalker or the person with a shameful family history, what the situation really demands is for the person to acknowledge that the harm-causing tendency in question is a part of who they are, even if they are not morally responsible for it, and to commit to preventing it from happening again. In fact, one might notice that a shared feature of expressivism-friendly cases is that the individual's blameworthiness appears to increase with each subsequent incident. While forgetting a birthday or mixing up names one time may be brushed off without indignation, once it becomes a pattern, it is difficult to overlook, even if the person is not conscious of what they are doing in each specific instance. The cumulative nature of these cases is nothing more than a consequence of tracing. The reason Smith would be more to blame the second or third time she forgets her friend's birthday is because there is a clear instance (immediately after the first time she forgets) in which she was likely conscious of the moral relevance and associated risks of not marking her calendar, and to which her later forgetting may be traced. As more such instances accrue, the scenario gradually shifts from one of pure ownership without responsibility to one of robust moral responsibility due to tracing.

A problem, however, does arise in cases of total obliviousness, such as that of Arthur and his microaggression. If Arthur is never made aware of what he has done, not during or after the incident, can he be considered responsible if he continues to mix up his Asian students repeatedly throughout the rest of the semester? I maintain that no, he cannot be held responsible, so long as he remains totally oblivious to his actions—*and* so long as his obliviousness is itself not under his awareness or otherwise traceable. But this scenario is far-fetched, as he would most likely eventually notice or be notified of one of his misdeeds, or at least become aware of his own general obliviousness towards his in-class behavior, at which point he would pick up some responsibility. Another consequence of this line of reasoning is the notion that we have a special power to *mark* individuals for responsibility for future lapses for which they otherwise would not be considered responsible. By making an individual aware of their own deficient priorities or morally perilous unconscious dispositions—for instance, by pointing out to Arthur that he mixes up his Asian students—you are creating an origin point that subsequent wrongs can then be traced back to, essentially putting the agent on the hook for their future expressivism-friendly wrongdoings.

To summarize, despite its theoretical superiority to expressivism, volitionalism seems to be undermined by a class of cases that elicit intuitions of moral responsibility, such as instances of forgetting, insensitivity, and microaggression. Rather than trying to vindicate these judgments of moral responsibility through tracing, we can debunk them by showing that they are more aptly captured as instances of ownership sans responsibility. This puts expressivism-friendly cases in the same class as involuntary, inherited, or unlucky acts of harm, wherein the relevant individual ought to acknowledge the harm and their fundamental connection to it, express dismay, and commit to preventing future occurrences, without being subjected to any moral appraisal.

With the relationship between consciousness and moral responsibility more clearly marked out, the only task remaining[6] is to apply this relationship to the central question of this thesis: Is conscious awareness of risk truly necessary for an agent to be considered reckless? Based on the

[6] It is also worth addressing what it actually means to *disregard* a consciously apprehended risk. Specifically, one might make a distinction between genuinely disregarding a risk and merely acting in spite of it, even if one is conscious of the risk in both cases. The difference is that in the former, one must actively think, *I know there is a risk but I am going to perform this action anyway*, while in the latter, one only needs to consciously know there is a risk and perform the action anyway. If this distinction is in fact a real one—which it may not be—I would argue that the former requirement is far too stringent, as very few ordinary, uncontroversial cases of recklessness involve the agent making such an explicit mental declaration. Instead, it is sufficient to require only that the agent be conscious of the risk of harm in question and perform the risky action nevertheless.

neuroscientific data and philosophical analysis, the answer approaches yes. First, the results of the EEG experiment provide preliminary evidence that there is a concrete difference in brain activity during choices involving conscious versus unconscious processing of risk, which supports the idea that conscious awareness of risk is a distinguishable feature of an agent's mental state. Next, the theoretical case for volitionalism, and rebuttal of expressivism-friendly intuitions, confirms that the distinction between conscious and unconscious processing of risk is morally relevant: If an agent is not consciously aware of the risk of harm of an action, and if disregarding a risk of harm provides a pro tanto reason to believe an action is wrong, then the agent cannot be consciously aware of the facts that give the action moral significance in this respect. As a result, they are not morally responsible for the reckless quality of their action, though they still may be answerable to questions of negligence.

Supplementary Materials:

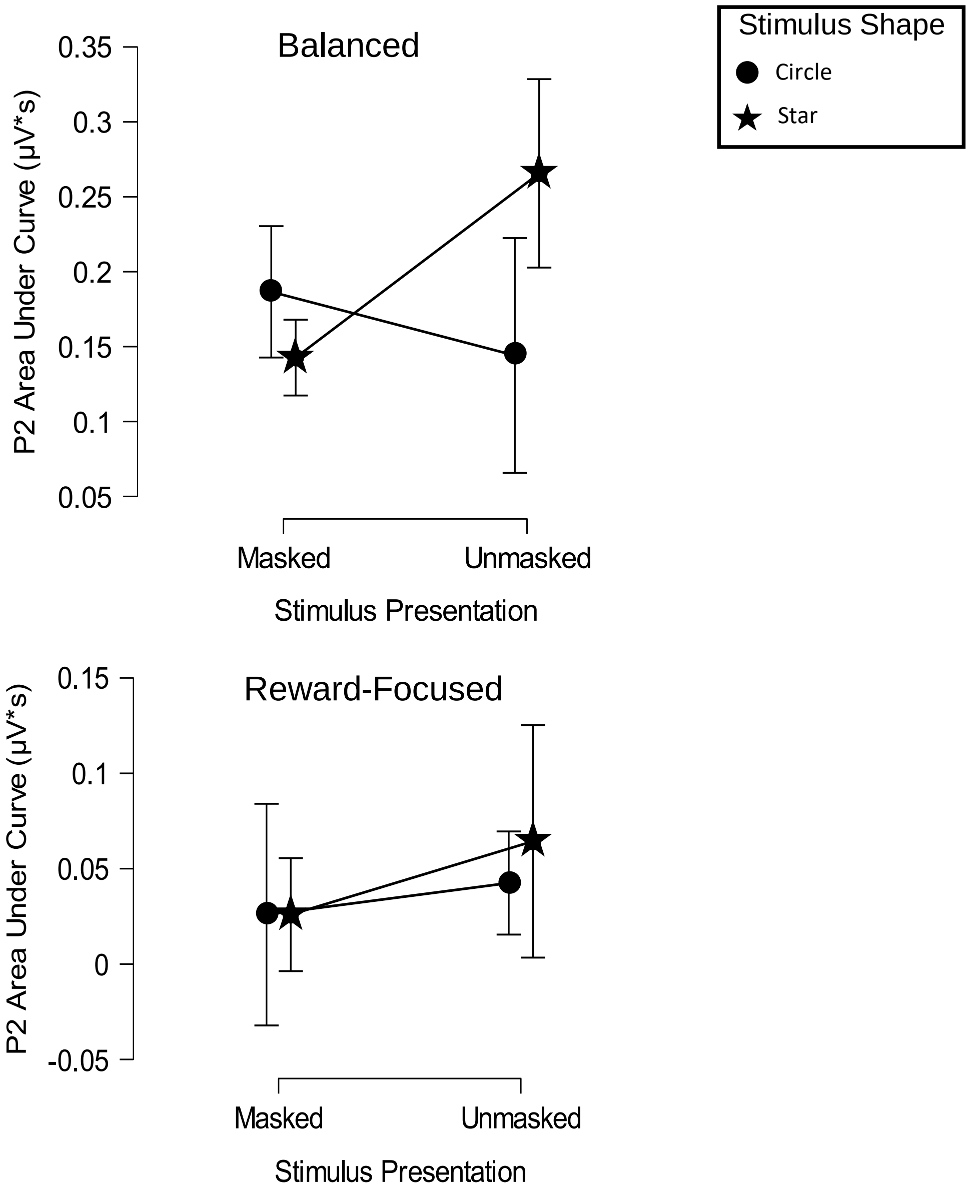


Figure S1 | Strategy Group Differences in P2 Amplitude with 95% Confidence Intervals. With participant strategy included as a between-subject effect, AUC analysis with a baseline interval of 600 ms prior to choice onset revealed a significant effect of stimulus shape ($p = .023$), participant strategy ($p = .039$), stimulus shape × stimulus presentation ($p = .005$), stimulus shape × participant strategy ($p = .011$), and shape × presentation × participant strategy ($p = .019$).

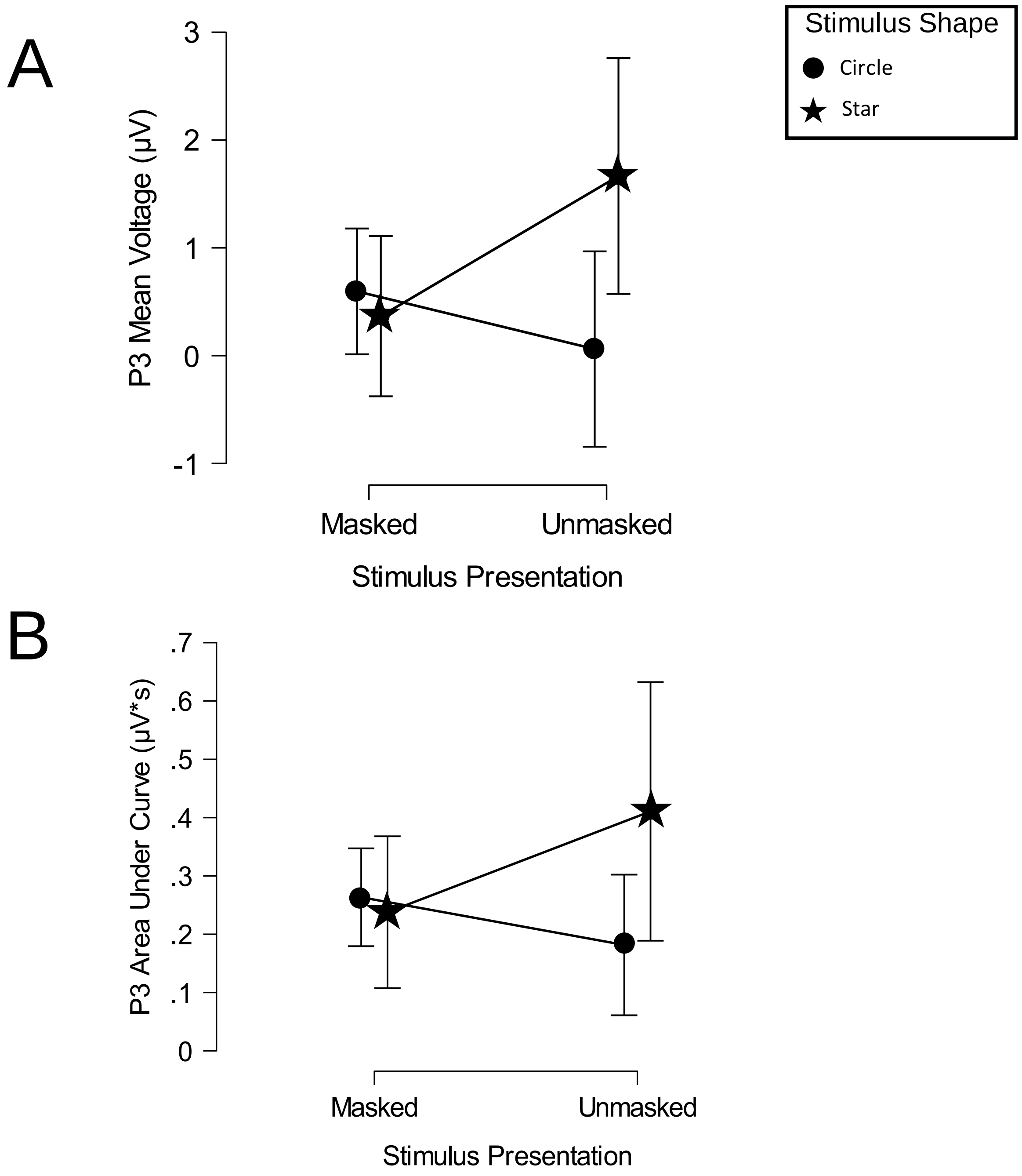
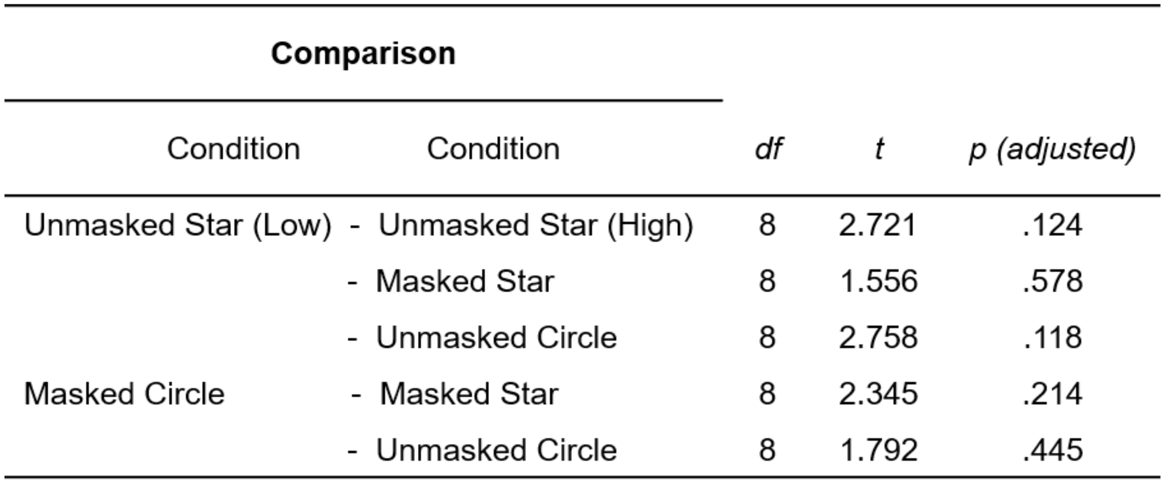
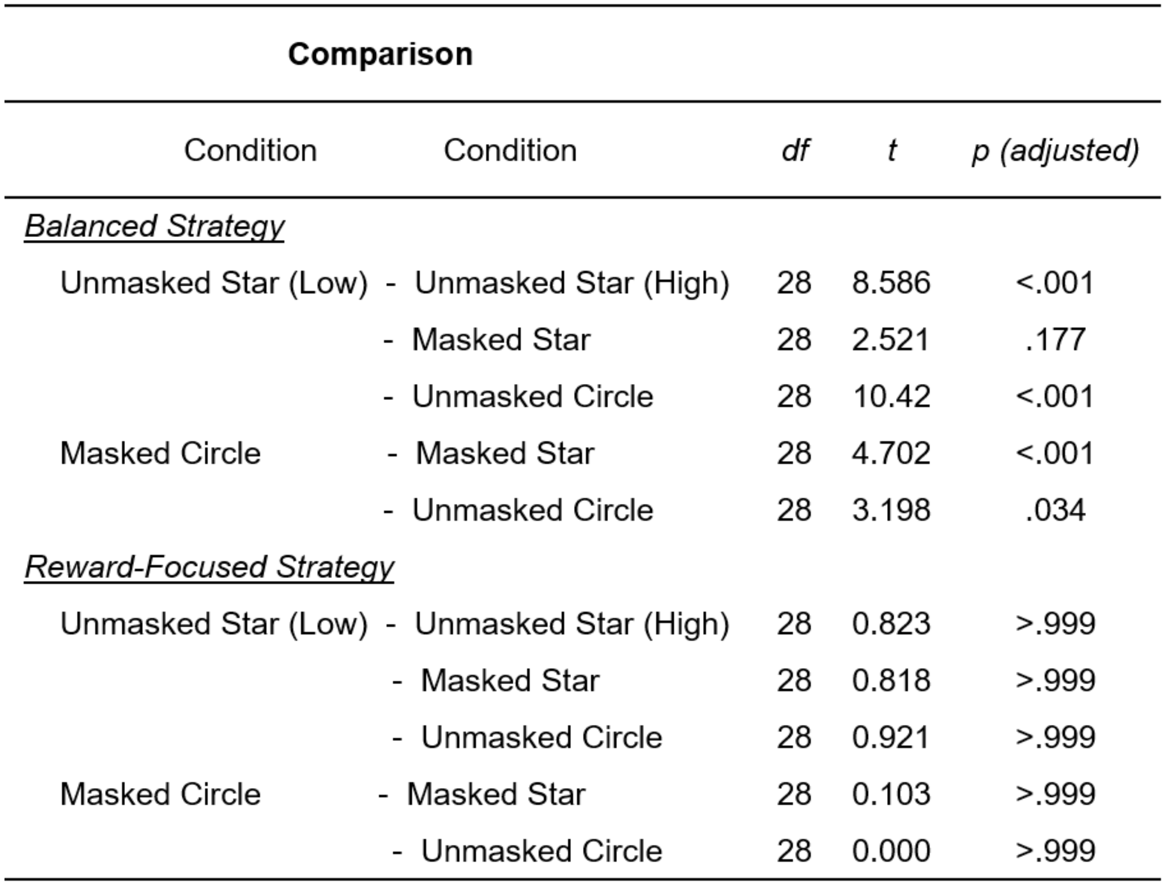


Figure S2 | P3 Amplitude Analysis in Low-Reward Trials. To assess whether P3 amplitude differences could be explained by differences in reward size, I performed a repeated measures ANOVA on only trials with low reward value. The error bars represent a confidence interval of 95%. **A.** The analysis of mean voltage showed a significant effect of stimulus shape (p = .010) and participant strategy (p = .005). **B.** The analysis of AUC with a baseline interval of 200 ms before choice onset showed a significant effect of stimulus shape (p = .037), and participant strategy (p = .009).



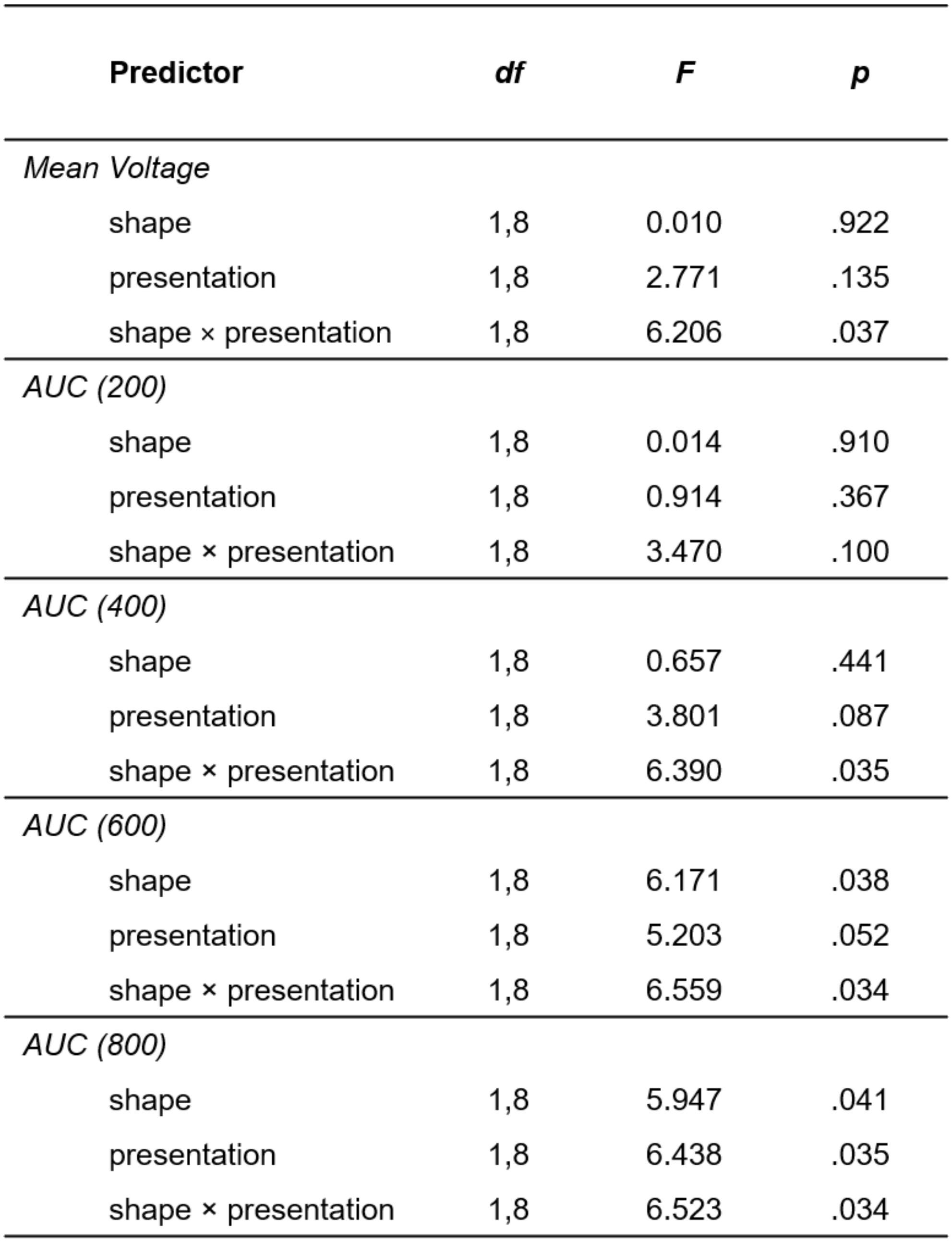
Comparison					
Condition		Condition	*df*	*t*	*p (adjusted)*
Unmasked Star (Low)	-	Unmasked Star (High)	8	2.721	.124
	-	Masked Star	8	1.556	.578
	-	Unmasked Circle	8	2.758	.118
Masked Circle	-	Masked Star	8	2.345	.214
	-	Unmasked Circle	8	1.792	.445

Table S1 | Post Hoc Comparisons of Reward-Selection Rates. Šidák correction applied.



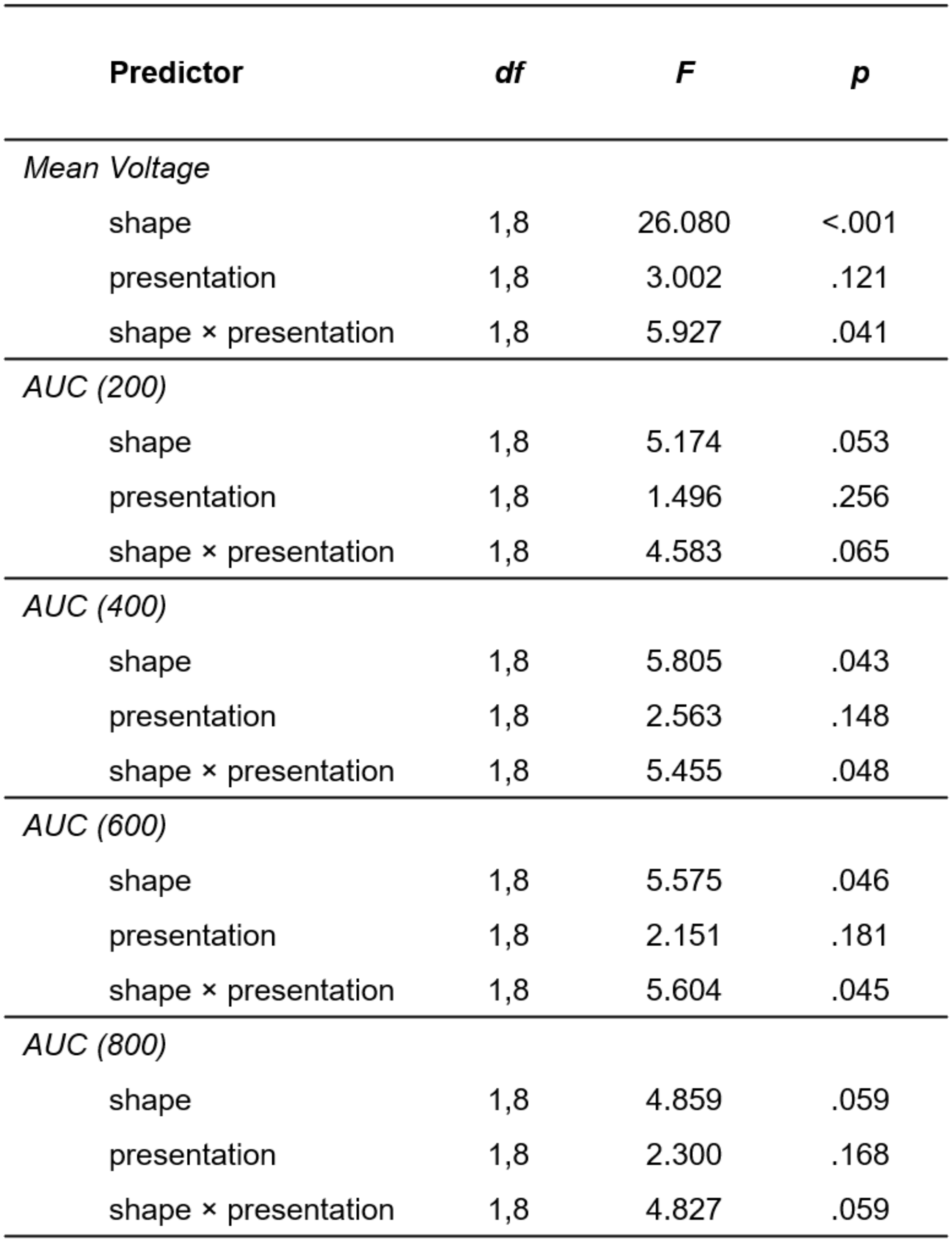
Comparison					
Condition		Condition	*df*	*t*	*p (adjusted)*
Balanced Strategy					
Unmasked Star (Low)	-	Unmasked Star (High)	28	8.586	<.001
	-	Masked Star	28	2.521	.177
	-	Unmasked Circle	28	10.42	<.001
Masked Circle	-	Masked Star	28	4.702	<.001
	-	Unmasked Circle	28	3.198	.034
Reward-Focused Strategy					
Unmasked Star (Low)	-	Unmasked Star (High)	28	0.823	>.999
	-	Masked Star	28	0.818	>.999
	-	Unmasked Circle	28	0.921	>.999
Masked Circle	-	Masked Star	28	0.103	>.999
	-	Unmasked Circle	28	0.000	>.999

Table S2 | Post Hoc Comparisons of Reward-Selection Rates within Strategy Groups. Bonferroni correction applied.



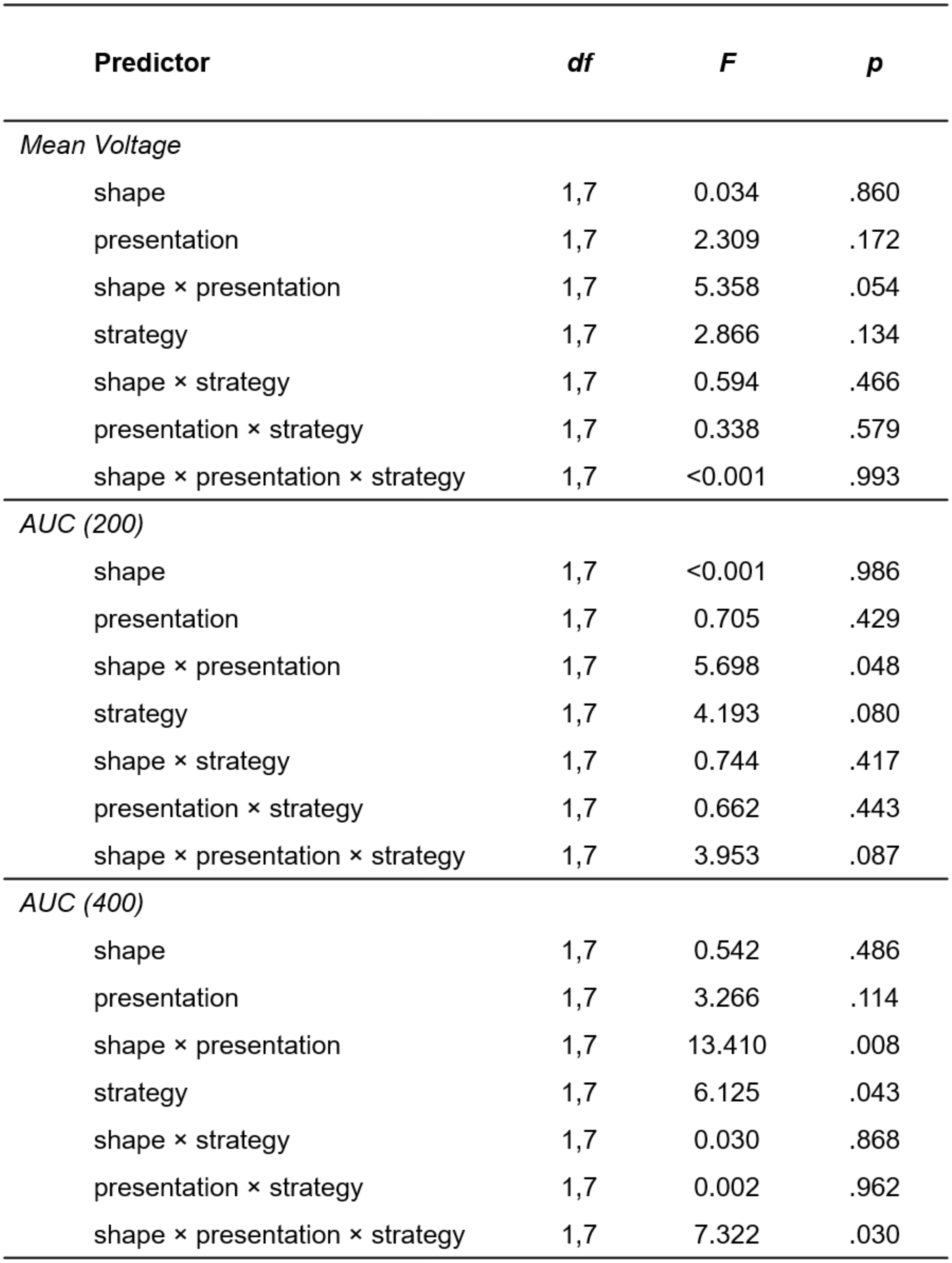
Predictor	*df*	*F*	*p*
Mean Voltage			
shape	1,8	0.010	.922
presentation	1,8	2.771	.135
shape × presentation	1,8	6.206	.037
AUC (200)			
shape	1,8	0.014	.910
presentation	1,8	0.914	.367
shape × presentation	1,8	3.470	.100
AUC (400)			
shape	1,8	0.657	.441
presentation	1,8	3.801	.087
shape × presentation	1,8	6.390	.035
AUC (600)			
shape	1,8	6.171	.038
presentation	1,8	5.203	.052
shape × presentation	1,8	6.559	.034
AUC (800)			
shape	1,8	5.947	.041
presentation	1,8	6.438	.035
shape × presentation	1,8	6.523	.034

Table S3 | P2 Amplitude ANOVA Tables.



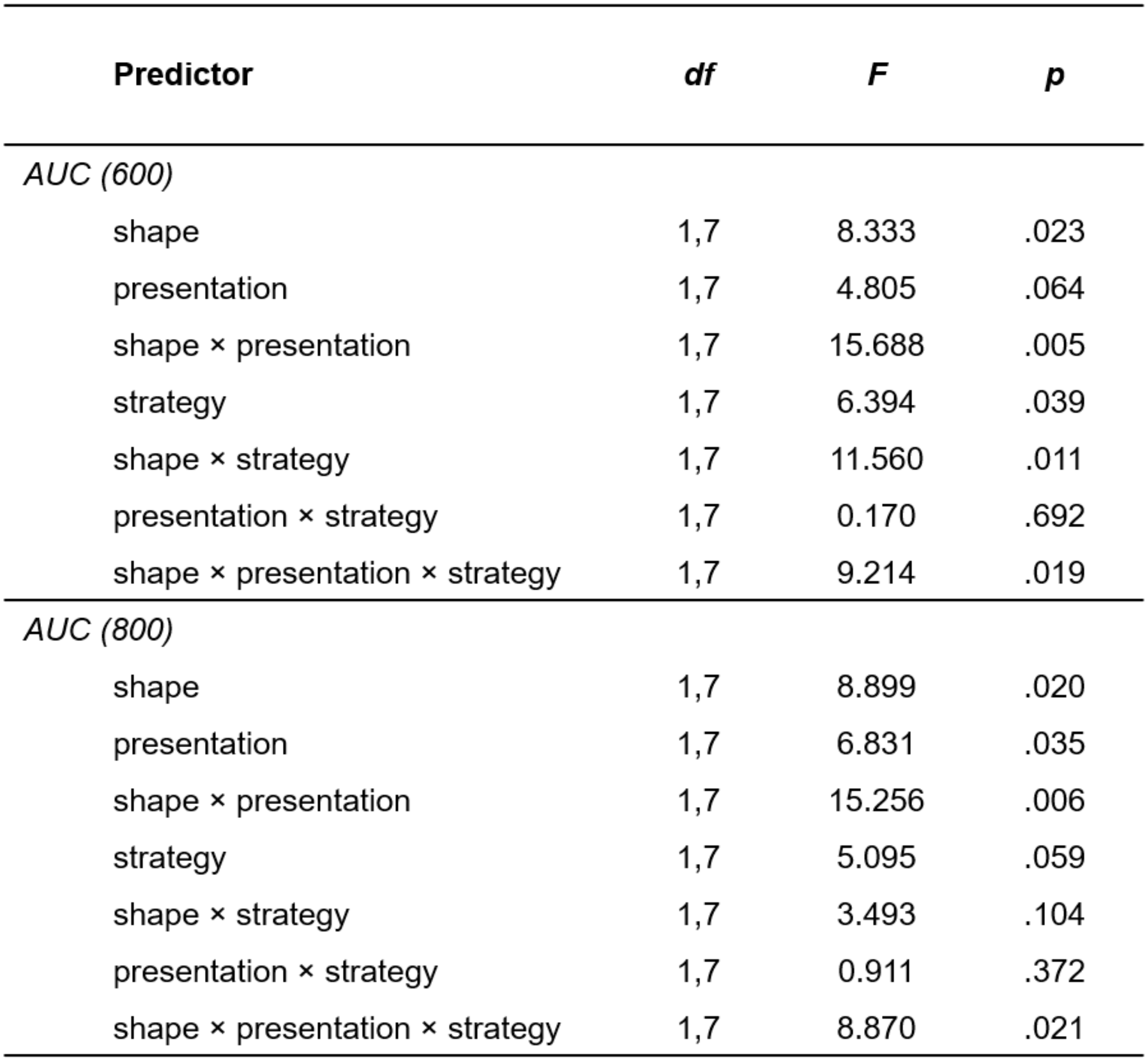
Predictor	*df*	*F*	*p*
Mean Voltage			
shape	1,8	26.080	<.001
presentation	1,8	3.002	.121
shape × presentation	1,8	5.927	.041
AUC (200)			
shape	1,8	5.174	.053
presentation	1,8	1.496	.256
shape × presentation	1,8	4.583	.065
AUC (400)			
shape	1,8	5.805	.043
presentation	1,8	2.563	.148
shape × presentation	1,8	5.455	.048
AUC (600)			
shape	1,8	5.575	.046
presentation	1,8	2.151	.181
shape × presentation	1,8	5.604	.045
AUC (800)			
shape	1,8	4.859	.059
presentation	1,8	2.300	.168
shape × presentation	1,8	4.827	.059

Table S4 | P3 Amplitude ANOVA Tables.



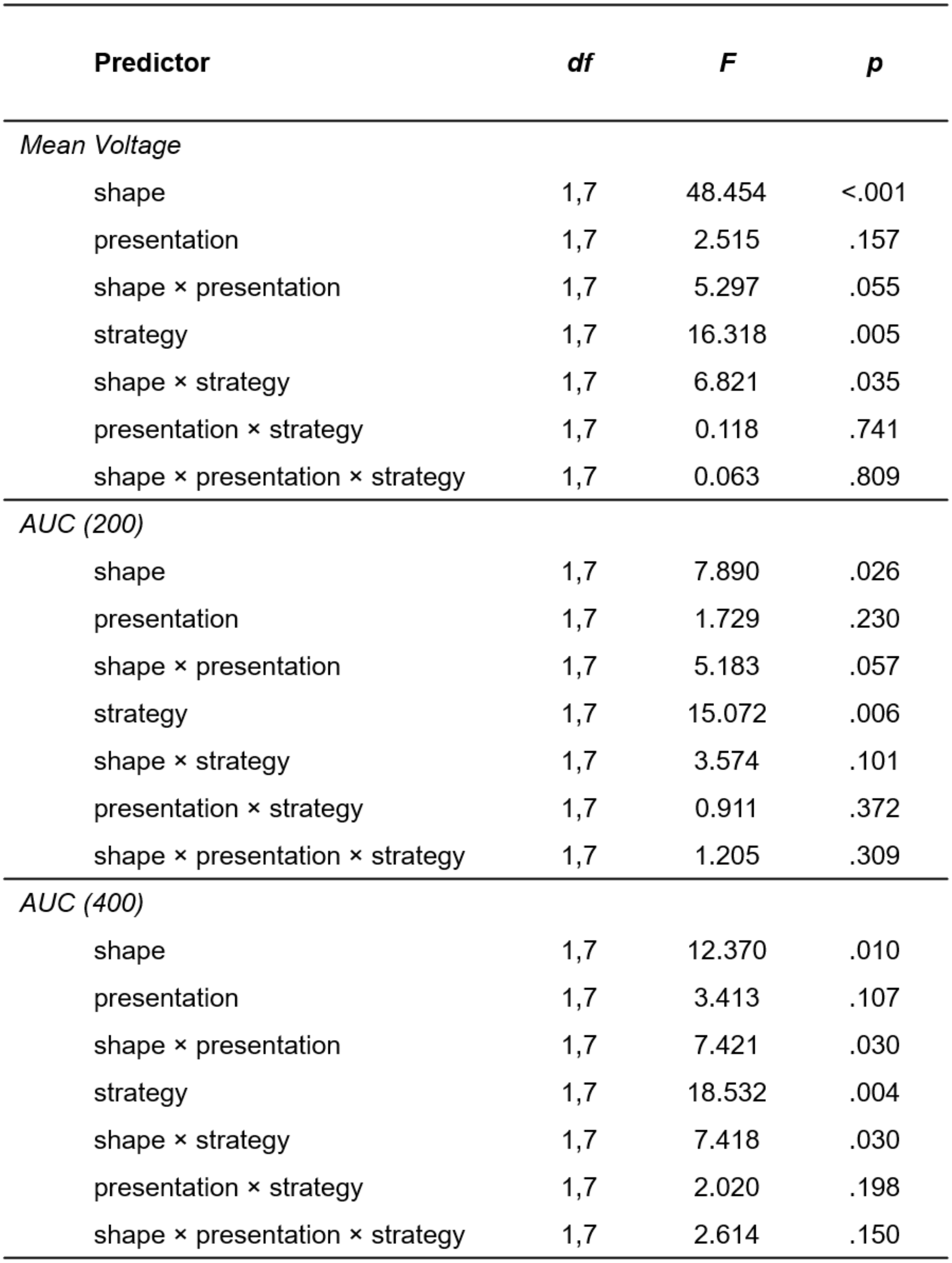
Predictor	*df*	*F*	*p*
Mean Voltage			
shape	1,7	0.034	.860
presentation	1,7	2.309	.172
shape × presentation	1,7	5.358	.054
strategy	1,7	2.866	.134
shape × strategy	1,7	0.594	.466
presentation × strategy	1,7	0.338	.579
shape × presentation × strategy	1,7	<0.001	.993
AUC (200)			
shape	1,7	<0.001	.986
presentation	1,7	0.705	.429
shape × presentation	1,7	5.698	.048
strategy	1,7	4.193	.080
shape × strategy	1,7	0.744	.417
presentation × strategy	1,7	0.662	.443
shape × presentation × strategy	1,7	3.953	.087
AUC (400)			
shape	1,7	0.542	.486
presentation	1,7	3.266	.114
shape × presentation	1,7	13.410	.008
strategy	1,7	6.125	.043
shape × strategy	1,7	0.030	.868
presentation × strategy	1,7	0.002	.962
shape × presentation × strategy	1,7	7.322	.030

Table S5 | P2 Amplitude ANOVA Tables with Strategy Included. Table continues onto next page.



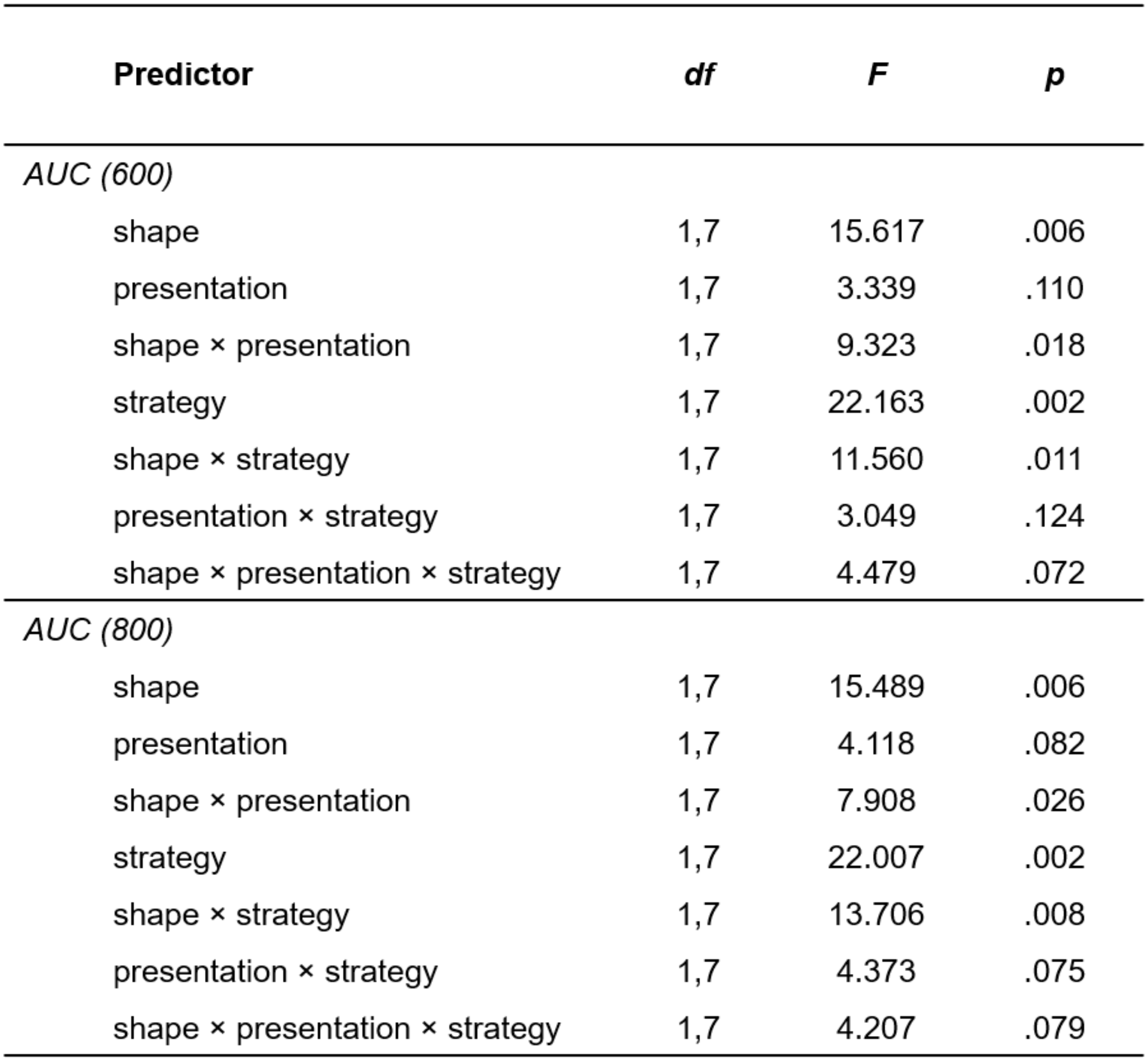
Predictor	*df*	*F*	*p*
AUC (600)			
shape	1,7	8.333	.023
presentation	1,7	4.805	.064
shape × presentation	1,7	15.688	.005
strategy	1,7	6.394	.039
shape × strategy	1,7	11.560	.011
presentation × strategy	1,7	0.170	.692
shape × presentation × strategy	1,7	9.214	.019
AUC (800)			
shape	1,7	8.899	.020
presentation	1,7	6.831	.035
shape × presentation	1,7	15.256	.006
strategy	1,7	5.095	.059
shape × strategy	1,7	3.493	.104
presentation × strategy	1,7	0.911	.372
shape × presentation × strategy	1,7	8.870	.021

Table S5 | P2 Amplitude ANOVA Tables with Strategy Included (cont.).



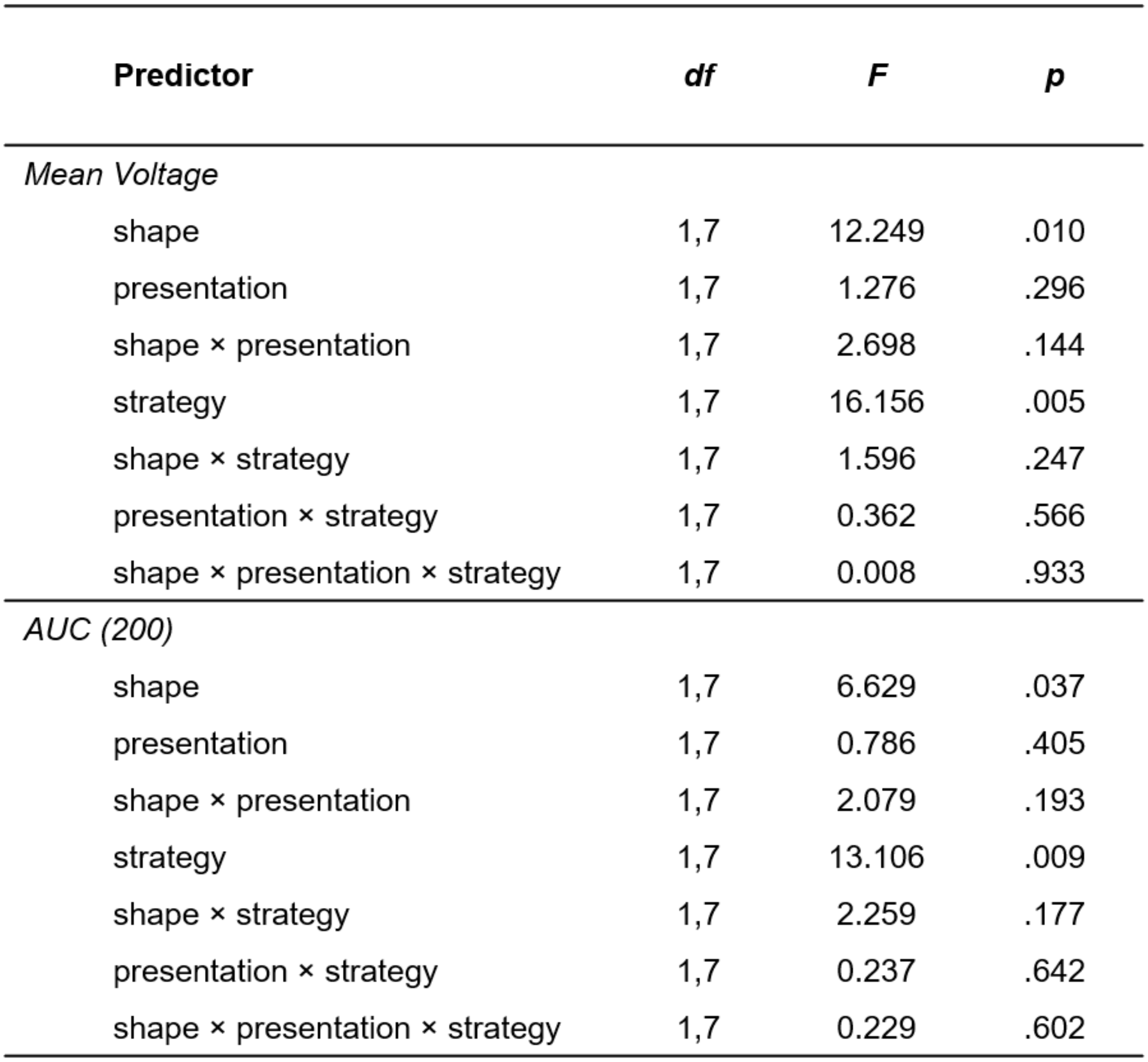
Predictor	*df*	*F*	*p*
Mean Voltage			
shape	1,7	48.454	<.001
presentation	1,7	2.515	.157
shape × presentation	1,7	5.297	.055
strategy	1,7	16.318	.005
shape × strategy	1,7	6.821	.035
presentation × strategy	1,7	0.118	.741
shape × presentation × strategy	1,7	0.063	.809
AUC (200)			
shape	1,7	7.890	.026
presentation	1,7	1.729	.230
shape × presentation	1,7	5.183	.057
strategy	1,7	15.072	.006
shape × strategy	1,7	3.574	.101
presentation × strategy	1,7	0.911	.372
shape × presentation × strategy	1,7	1.205	.309
AUC (400)			
shape	1,7	12.370	.010
presentation	1,7	3.413	.107
shape × presentation	1,7	7.421	.030
strategy	1,7	18.532	.004
shape × strategy	1,7	7.418	.030
presentation × strategy	1,7	2.020	.198
shape × presentation × strategy	1,7	2.614	.150

Table S6 | P3 Amplitude ANOVA Tables with Strategy Included. Table continues onto next page.



Predictor	*df*	*F*	*p*
AUC (600)			
shape	1,7	15.617	.006
presentation	1,7	3.339	.110
shape × presentation	1,7	9.323	.018
strategy	1,7	22.163	.002
shape × strategy	1,7	11.560	.011
presentation × strategy	1,7	3.049	.124
shape × presentation × strategy	1,7	4.479	.072
AUC (800)			
shape	1,7	15.489	.006
presentation	1,7	4.118	.082
shape × presentation	1,7	7.908	.026
strategy	1,7	22.007	.002
shape × strategy	1,7	13.706	.008
presentation × strategy	1,7	4.373	.075
shape × presentation × strategy	1,7	4.207	.079

Table S6 | P3 Amplitude ANOVA Tables with Strategy Included (cont.).



Predictor	*df*	*F*	*p*
Mean Voltage			
shape	1,7	12.249	.010
presentation	1,7	1.276	.296
shape × presentation	1,7	2.698	.144
strategy	1,7	16.156	.005
shape × strategy	1,7	1.596	.247
presentation × strategy	1,7	0.362	.566
shape × presentation × strategy	1,7	0.008	.933
AUC (200)			
shape	1,7	6.629	.037
presentation	1,7	0.786	.405
shape × presentation	1,7	2.079	.193
strategy	1,7	13.106	.009
shape × strategy	1,7	2.259	.177
presentation × strategy	1,7	0.237	.642
shape × presentation × strategy	1,7	0.229	.602

Table S7 | P3 Amplitude ANOVA Tables with Low Reward Only.

References:

Baumeister RF, Masicampo EJ (2010) Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal–culture interface. Psychological Review 117:945–971.

Breitmeyer BG, Tapia E, Kafalıgönül H, Öğmen H (2008) Metacontrast masking and stimulus contrast polarity. Vision Research 48:2433–2438.

Chalmers DJ (1996) The conscious mind: in search of a fundamental theory. New York: Oxford University Press.

Chen P, Qiu J, Li H, Zhang Q (2009) Spatiotemporal cortical activation underlying dilemma decision-making: An event-related potential study. Biological Psychology 82:111–115.

Clayson PE, Carbine KA, Baldwin SA, Larson MJ (2019) Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. Psychophysiology 56 Available at: https://onlinelibrary.wiley.com/doi/10.1111/psyp.13437 [Accessed March 8, 2023].

Commonwealth of Massachusetts (n.d.) General Law - Part IV, Title I, Chapter 265, Section 13L. Available at: https://malegislature.gov/Laws/GeneralLaws/PartIV/TitleI/Chapter265/Section13L#:~:text=Whoever%20wantonly%20or%20recklessly%20engages,of%20correction%20for%20not%20more.

Cornsweet TN (1962) The Staircase-Method in Psychophysics. The American Journal of Psychology 75:485.

De Renzi E (2000) Disorders of Visual Recognition. Semin Neurol 20:479–486.

Dehaene S (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79:1–37.

Dehaene S, Naccache L, Cohen L, Bihan DL, Mangin J-F, Poline J-B, Rivière D (2001) Cerebral mechanisms of word masking and unconscious repetition priming. Nat Neurosci 4:752–758.

DeWall CN, Baumeister RF, Masicampo EJ (2008) Evidence that logical reasoning depends on conscious processing. Consciousness and Cognition 17:628–645.

Donchin E, Coles MGH (1988) Is the P300 component a manifestation of context updating? Behav Brain Sci 11:357.

Farah MJ, Rabinowitz C (2003) GENETIC AND ENVIRONMENTAL INFLUENCES ON THE ORGANISATION OF SEMANTIC MEMORY IN THE BRAIN:IS “LIVING THINGS” AN INNATE CATEGORY? Cognitive Neuropsychology 20:401–408.

Fischer JM, Ravizza M (2000) Responsibility and control: a theory of moral responsibility, First paperback ed. Cambridge: Cambridge University Press.

Fischer JM, Tognazzini NA (2009) The Truth about Tracing. Noûs 43:531–556.

Fries P, Roelfsema PR, Engel AK, König P, Singer W (1997) Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. Proc Natl Acad Sci USA 94:12699–12704.

Gaillard R, Dehaene S, Adam C, Clémenceau S, Hasboun D, Baulac M, Cohen L, Naccache L (2009) Converging Intracranial Markers of Conscious Access Ungerleider L, ed. PLoS Biol 7:e1000061.

Garrigan B, Adlam ALR, Langdon PE (2016) The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements. Brain and Cognition 108:88–97.

Goldstein RZ, Cottone LA, Jia Z, Maloney T, Volkow ND, Squires NK (2006) The effect of graded monetary reward on cognitive event-related potentials and behavior in young healthy adults. International Journal of Psychophysiology 62:272–279.

Goodale MA, Milner AD, Jakobson LS, Carey DP (1991) A neurological dissociation between perceiving objects and grasping them. Nature 349:154–156.

Greenawalt RK (1991) Model penal code and commentaries (Official draft and revised commentaries), with text of the model penal code as adopted at the 1962 annual meeting of the American Law Institute at Washington, D.C., May 24, 1962. Philadelphia: American Law Institute.

Greenwald AG, Krieger LH (2006) Implicit Bias: Scientific Foundations. California Law Review 94:945.

Gregoriou GG, Gotts SJ, Zhou H, Desimone R (2009) High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention. Science 324:1207–1210.

Halgren E, Baudena P, Clarke JM, Heit G, Liégeois C, Chauvel P, Musolino A (1995) Intracerebral potentials to rare target and distractor auditory and visual stimuli. I. Superior temporal plane and parietal lobe. Electroencephalography and Clinical Neurophysiology 94:191–220.

Hasson U, Glucksberg S (2006) Does understanding negation entail affirmation? Journal of Pragmatics 38:1015–1032.

Hu X, Mai X (2021) Social value orientation modulates fairness processing during social decision-making: evidence from behavior and brain potentials. Social Cognitive and Affective Neuroscience 16:670–682.

Im C-H (2018) Basics of EEG: Generation, Acquisition, and Applications of EEG. In: Computational EEG Analysis (Im C-H, ed), pp 3–11 Biological and Medical Physics, Biomedical Engineering. Singapore: Springer Singapore. Available at: http://link.springer.com/10.1007/978-981-13-0908-3_1 [Accessed February 22, 2023].

Key APF, Dove GO, Maguire MJ (2005) Linking Brainwaves to the Brain: An ERP Primer. Developmental Neuropsychology 27:183–215.

Knight RT (1984) Decreased response to novel stimuli after prefrontal lesions in man. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 59:9–20.

Laureys S, Faymonville ME, Peigneux P, Damas P, Lambermont B, Del Fiore G, Degueldre C, Aerts J, Luxen A, Franck G, Lamy M, Moonen G, Maquet P (2002) Cortical processing of noxious somatosensory stimuli in the persistent vegetative state. Neuroimage 17:732–741.

Laureys S, Lemaire C, Maquet P, Phillips C, Franck G (1999) Cerebral metabolism during vegetative state and after recovery to consciousness. Journal of Neurology, Neurosurgery & Psychiatry 67:121–122.

Legal Information Institute (n.d.) Mens Rea. Wex Legal Encyclopedia Available at: https://www.law.cornell.edu/wex/mens_rea.

Levy N (2014) Consciousness and moral responsibility, 1st ed. Oxford ; New York: Oxford University Press.

Light GA, Williams LE, Minow F, Sprock J, Rissling A, Sharp R, Swerdlow NR, Braff DL (2010) Electroencephalography (EEG) and Event-Related Potentials (ERPs) with Human Participants. CP Neuroscience 52 Available at: https://onlinelibrary.wiley.com/doi/10.1002/0471142301.ns0625s52 [Accessed February 24, 2023].

Lindholm E, Koriath JJ (1985) Analysis of multiple event related potential components in a tone discrimination task. International Journal of Psychophysiology 3:121–129.

Luck SJ (2014) An introduction to the event-related potential technique, Second edition. Cambridge, Massachusetts: The MIT Press.

Mahon BZ, Caramazza A (2009) Concepts and Categories: A Cognitive Neuropsychological Perspective. Annu Rev Psychol 60:27–51.

Maoz U, Yaffe G (2016) What does recent neuroscience tell us about criminal responsibility? J Law and the BioSci 3:120–139.

Mastropasqua T, Turatto M (2015) Attention is necessary for subliminal instrumental conditioning. Sci Rep 5:12920.

Melloni L, Molina C, Pena M, Torres D, Singer W, Rodriguez E (2007) Synchronization of Neural Activity across Cortical Areas Correlates with Conscious Perception. Journal of Neuroscience 27:2858–2865.

Mudrik L, Deouell LY (2022) Neuroscientific Evidence for Processing Without Awareness. Annu Rev Neurosci 45:403–423.

Nagel T (1974) What Is It Like to Be a Bat? The Philosophical Review 83:435.

Neys WD (2006) Dual Processing in Reasoning: Two Systems but One Reasoner. Psychol Sci 17:428–433.

Nieuwenhuis S, Aston-Jones G, Cohen JD (2005) Decision making, the P3, and the locus coeruleus--norepinephrine system. Psychological Bulletin 131:510–532.

Oberbroeckling LA (2021) Numerical Integration. In: Programming Mathematics Using MATLAB®, pp 183–191. Elsevier. Available at: https://linkinghub.elsevier.com/retrieve/pii/B978012817799000017X [Accessed February 23, 2023].

Office of the Law Revision Counsel (1956) 10 U.S. Code § 914 - Art. 114. Endangerment Offenses. Available at: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title10-section914&num=0&edition=prelim.

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK (2019) PsychoPy2: Experiments in behavior made easy. Behav Res 51:195–203.

Pessiglione M, Petrovic P, Daunizeau J, Palminteri S, Dolan RJ, Frith CD (2008) Subliminal Instrumental Conditioning Demonstrated in the Human Brain. Neuron 59:561–567.

Petrides M (2005) Lateral prefrontal cortex: architectonic and functional organization. Phil Trans R Soc B 360:781–795.

Polich J (2007) Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology 118:2128–2148.

Polich J (2011) Neuropsychology of P300. Oxford University Press. Available at: https://academic.oup.com/edited-volume/34558/chapter/293241580 [Accessed December 1, 2022].

Potts GF (2004) An ERP index of task relevance evaluation of visual stimuli. Brain and Cognition 56:5–13.

Schmälzle R (2008) Inuitive Risk Perception: A Neuroscientific Approach. Available at: https://kops.uni-konstanz.de/bitstream/handle/123456789/10085/Diss_Schmaelzle.pdf?sequence=1&isAllowed=y.

Smith AM (2005) Responsibility for Attitudes: Activity and Passivity in Mental Life. Ethics 115:236–271.

Smith AM (2008) Control, responsibility, and moral assessment. Philos Stud 138:367–392.

Srinivasan R, Russell DP, Edelman GM, Tononi G (1999) Increased Synchronization of Neuromagnetic Responses during Conscious Perception. J Neurosci 19:5435–5448.

Sur S, Sinha V (2009) Event-related potential: An overview. Ind Psychiatry J 18:70.

van Gaal S, Ridderinkhof KR, Scholte HS, Lamme VAF (2010) Unconscious Activation of the Prefrontal No-Go Network. Journal of Neuroscience 30:4143–4150.

Veen V van, Carter CS (2002) The Timing of Action-Monitoring Processes in the Anterior Cingulate Cortex. Journal of Cognitive Neuroscience 14:593–602.

Verbaarschot C, Farquhar J, Haselager P (2015) Lost in time... Consciousness and Cognition 33:300–315.

Vilares I, Wesley MJ, Ahn W-Y, Bonnie RJ, Hoffman M, Jones OD, Morse SJ, Yaffe G, Lohrenz T, Montague PR (2017) Predicting the knowledge–recklessness distinction in the human brain. Proc Natl Acad Sci USA 114:3222–3227.

Williams BAO, Nagel T (1976) Moral Luck. Aristot Soc Suppl Vol 50:115–152.

Yoder KJ, Decety J (2018) The neuroscience of morality and social decision-making. Psychology, Crime & Law 24:279–295.

Yoo CY (2008) Unconscious processing of Web advertising: Effects on implicit memory, attitude toward the brand, and consideration set. Journal of Interactive Marketing 22:2–18.

Zhan Y, Xiao X, Tan Q, Li J, Fan W, Chen J, Zhong Y (2020) Neural correlations of the influence of self-relevance on moral decision-making involving a trade-off between harm and reward. Psychophysiology 57 Available at: https://onlinelibrary.wiley.com/doi/10.1111/psyp.13590 [Accessed December 2, 2022].