Neural correlates of three putative mechanisms for communication-driven behavior change

By

Matt Minich

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Mass Communications)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 08/23/2023

This dissertation is approved by the following members of the Final Oral Committee:
Christopher N. Cascio, Assistant Professor, Journalism and Mass Communication
Karyn Riddle, Robert Taylor Professor, Journalism and Mass Communication
Markus Brauer, Professor, Psychology
John Rudolph, Professor, Curriculum and Instruction
Dhavan Shah, Maier-Bascom Professor, Journalism and Mass Communication
Ralf Schmälzle, Associate Professor, Communication Science, Michigan State University

ACKNOWLEDGEMENTS

To my great relief, this dissertation was not undertaken alone—I have enjoyed the support and guidance of many throughout the five years of my PhD journey. Specifically, I owe thanks to my advisor Dr. Christopher N. Cascio for his guidance, feedback, and support during the development of this dissertation and throughout my doctoral studies. I would also like to thank the many people who collaborated on the research presented in this dissertation, particularly Chen-Ting Chang, Arina Tveleneva, and Lauren Kriss. These three people were integral to the design and execution of the studies presented here. For their part in the analysis and writing of chapter four, I thank Lynne Cotter and Dr. Sijia Yang. For their patience and support, I'm indebted to Mark and Debra Minich, Kaitlyn and Nathan Beekman, my partner Katrina Virta, and my cat Cheeto.

ABSTRACT

NEURAL CORRELATES OF THREE PUTATIVE MECHANISMS FOR COMMUNICATION-DRIVEN BEHAVIOR CHANGE

Matt Minich

Theories of persuasion are integral to the study of mass communication, and these theories are often used to inform the design of mass media campaigns promoting public health. Such campaigns are often effective at changing population-level health behaviors, but it is not clear that the message design features advocated by theories of persuasion are meaningful factors in campaign success. To better gauge the real-world importance of persuasion theory, therefore, it is important to achieve a better understanding of the mechanisms by which these design features are thought to contribute to attitude or behavior change. The current dissertation aims to improve that understanding by examining the ways theory-inspired message design features affect activity in the brain during message encoding. Specifically, we examined the effects of three features. First, we sought to build on past findings that the gain/loss framing of persuasive messages affected activation of a brain region associated with self-relevance and subjective value. Next, we used an inter-subject correlation procedure and an online survey to test a set of hypothesized neural correlates of psychological reactance in response to a set of public service announcements about driving under the influence of cannabis. Finally, we tested whether placing pictorial warning labels on social media posts promoting cannabis affected intentions to share those posts and activity in brain regions associated with decisions to share content on social media. Taken together, our findings suggest that neuroimaging methods can be used not only to uncover the processes that underlie persuasion, but also the ways those processes are affected by theoretically-inspired changes in message design

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND OVERVIEW

Introduction

The idea that theories from psychology, rhetoric, and other disciplines can inform the development of persuasive messages has been central to the study of mass media for more than a century (Briñol & Petty, 2012), preceding the emergence of Mass Communication as an academic discipline by several decades (Eadie, 2021). From early work like Carl Hovland's learning theory of persuasion (Hovland, Janis, & Kelley, 1953) to more contemporary theories like the revised Extended Parallel Process Model (So, 2013), would-be persuaders have long sought to use theories about the minds of individuals to change the minds of the masses. This is particularly important in the field of health communication, where persuasive messages are the primary tool used to improve public health. Theory-informed mass media campaigns reach large audiences through established channels like television, social media, and print, and they have been shown to drive both the adoption of healthy behaviors and the cessation of unhealthy behaviors (Wakefield, Loken & Hornik, 2010).

The success of mass media campaigns is not guaranteed, however—even well-funded campaigns sometimes fail to affect behavior change or can generate "boomerang effects" that drive attitudes and behaviors in the opposite direction as intended (Derzon & Lipsey, 2001). Outcomes are known to depend partly on message design (Noar, 2006), but scholars have struggled to identify design choices that are reliably associated with campaign success (O'Keefe & Hoeken, 2021). These difficulties might arise from an incomplete understanding of how different message features influence the mechanisms underlying persuasive message processing (Hagger et al., 2020), which is an area of active interest for public health agencies (Riddle &

Ferrer, 2015).

One way to improve our understanding of the psychological and affective mechanisms underlying persuasion is to observe how those mechanisms play out in the brain. Neuroimaging methods allow researchers to observe individuals' responses to stimuli in real time and without the need for introspection, limiting social desirability and recall biases that have long bedeviled the study of social processes (Lieberman, 2010). Neural processes such as those associated with subjective valuation, self, and social processing have been shown to be important components of persuasion (Falk & Scholz, 2018)—increased activity in regions associated with these processes during encoding of persuasive messages has been associated with behavior change in both small samples (Falk et al., 2010) and at the population level (Falk, Berkman & Lieberman, 2012). However, much of this research has focused on neural activity during aggregate level message exposure, making it difficult to determine how specific message features influence the underlying processes associated with persuasion.

The current dissertation aims to address this gap in the literature by examining how message features influence the underlying mechanisms associated with persuasion. Specifically, this dissertation aims to: 1) examine whether neural activity differs based on gain versus loss framed messages and how these differences relate to prospect theory; 2) examine whether neural activity differs based on fear versus humor appeals and how these differences relate to psychological reactance theory; and 3) examine whether neural activity differs based on messages with warning labels versus no warning labels and how these differences relate to a value-based model of information virality. Finally, this dissertation will aim to situate these findings within the larger context of persuasion theory and explain their implications for the

application of theory to message design. The importance and feasibility of these aims are outlined in the paragraphs below, followed by a brief overview of the remaining chapters.

Importance of understanding mechanisms of communication-driven behavior change

Theory-informed mass media campaigns have proven effective at changing health behaviors across a wide variety of contexts. Analyzing a set of nine meta-analyses of campaign effects, Snyder (2007) estimated that the average behavior change campaign has an effect of a full five percentage points. If the behavior in question was practiced by 20% of the population before distribution of a campaign, one can reasonably expect 25% to practice that behavior at the conclusion of the campaign. Notably, this number was estimated for campaigns that were not paired with any coercive element, such as regulations that change the prices or reduce the availability of given products. After a different review of 10 years of literature on the topic (some of it overlapping with Snyder, 2007), Noar (2006) also concluded that mass media campaigns usually have small to moderate effects on population-level health behaviors. A review published in the medical journal *Lancet* (Wakefield, Loken, & Hornik, 2010) reached a similar conclusion, finding that mass media campaigns could both increase positive behaviors and decrease negative behaviors across a variety of contexts.

Although research supports the effectiveness of mass media campaigns in general, it is also the case that some campaigns succeed in changing behavior while others fail. Snyder (2007) found that campaign success depended largely on the behavior being promoted, with campaigns promoting seatbelt use and proper dental care being the most effective and those designed to reduce youth cannabis use being the least. Wakefield, Loken and Hornik (2010) concluded in their review that campaigns were least likely to succeed when they advocated for habitual practices (e.g. sunscreen use or healthy eating), and most likely to succeed when advocating for

one-time behaviors (e.g. vaccinations or cancer screening). In addition to these factors, Noar (2006) concluded that campaigns were more likely to be effective when they were constructed and disseminated according to established best-practice principles such as message targeting, intentional channel selection, and the use of theory.

It has long been assumed that the design of mass media campaigns should be informed by theories of behavior change (Rice & Atkin, 2009; Randolph & Viswanath, 2004), but the real-world value of this practice is unclear. After a meta-analysis of 1,149 studies of mass media campaigns that tested 30 message features, O'Keefe & Hoeken (2021) concluded that popular theory-driven design choices (e.g. the use of narratives, gain/loss framing) exerted only small and unreliable effects on messages' persuasive success. Similar concerns have emerged regarding behavior change interventions as a whole (i.e. including non mass-media interventions), with some scholars arguing that though behavioral science can be used to predict behavior and attitude change in controlled studies, these observations ultimately reflect individual-level volatility that is washed out in the real world (Hagger & Weed, 2019). In response to these concerns, some have called for a more scientific approach of the study of behavior change, in which studies focus not only on the interactions between intervention features and outcomes but also on uncovering the putative mechanisms by which those features are thought to have their effects (Hagger et al., 2020; Aklin et al; 2020).

In sum, current research into the effects of mass media behavior change campaigns suggests an incomplete understanding of the mechanisms thought to drive campaign success. Though research suggests media campaigns can be effective (Snyder et al., 2007) and that success of these campaigns depends at least partly on message design (Noar, 2006), recent evidence suggests that theory-driven design choices do not have strong or reliable effects on that

success (O'Keefe & Hoeken, 2021). Similar findings in adjacent fields have been met with newfound attention to the putative mechanisms driving the effects of intervention design on behavior change outcomes (Hagger et al., 2020). Thus, more research is needed into the ways theory-driven design choices in mass media campaigns impact the cognitive and affective processes that are theorized to results in campaign success.

Feasibility of understanding message effects mechanisms through neuroimaging

Theories of persuasion often involve complex cognitive and affective processes such as elaboration (Petty & Cacioppo, 1981), narrative transportation (Green & Brock, 2000), and reactance (Brehm, 1966). These processes are assumed to be essential, usually causal antecedents to behavior change, but they are by nature difficult to measure. Researchers most often capture these processes with self-report measures, which are vulnerable to several biases (Paulhus & Vazire, 2007) and demand time and cognitive energy for participant introspection. Self-report measures are also limited because they cannot assess these processes in the moment they occur, which complicates inferences about their place in theorized causal chains. The use of neuroimaging methods like functional magnetic resonance imaging (fMRI) can address some of these concerns because these methods allow researchers to observe processes in real time and without need for participant introspection. This approach has proven to be a helpful supplement to self-report in the study of media effects, providing useful insights for a range of theoretical discussions (Weber et al., 2018).

Regarding the effects of mass media campaigns, neuroimaging research suggests that successful persuasion can be predicted by patterns of neural activity observed during the encoding of (i.e. reading, watching, or listening to) persuasive messages. Specifically, media messages appear to be more effective when they induce activity in brain regions associated with

processing self-relevance and value: the ventromedial prefrontal cortex (vmPFC), and ventral striatum (VS).

When Falk et al. (2010) presented participants undergoing fMRI with information about the benefits of sunscreen use, they found that participants who exhibited more activation within the vmPFC during message encoding were also more likely to report having used sunscreen the week after the scan. Notably, they found that vmPFC activation predicted variance in sunscreen use over and above that predicted by self-reported intentions. Similarly, Chua et al. (2011) found that smokers who exhibited more vmPFC activity while viewing tobacco cessation messages were more likely to have abstained from smoking four full months after their scanning session. Activity in the vmPFC has also proven successful in making *message-level* persuasion predictions. For example, when Falk et al. (2012) showed three sets of anti-smoking PSAs to participants undergoing fMRI, they found that a rank-ordering of the campaigns by average vmPFC activation aligned with a rank-ordering of the increases in call volume to a tobacco quitline in the months after each campaign was broadcast (driving smokers to call the quitline was the intent of all campaigns).

Activation of the vmPFC during message encoding is often accompanied by activation of the VS. For example, Cooper et al. (2018) found that functional connectivity (a measure of the degree to which networks interact) between the vmPFC and VS during the encoding of anti-smoking messages by smokers was associated with reductions in smoking after the scanning appointment. These regions are thought to comprise a network that is associated with making judgments of value and self-relevance, a process known as *subjective valuation* (Falk & Scholz, 2018) or *self-value integration* (Vezich et al., 2017). During encoding of persuasive messages, this process is thought to involve both recognition of value within message content and

estimations of the value of incorporating the advocated attitude or behavior changes into self-concept.

Importantly, self-relevance and subjective value are not the only neural processes that have been associated with persuasion. Social processes like mentalizing (inference about the thoughts or intentions of others) are also thought to play a role, particularly in the contexts of social influence (Cascio, Scholz, & Falk, 2015) and information sharing (Baek et al., 2017). Some work has also investigated processes like counterarguing (Liu et al., 2021; Weber et al., 2015), and emotion regulation (Doré et al., 2019). These processes may feed into the overall value calculations thought to inform persuasion and understanding them may be key to a better understanding of the mechanisms by which message design choices cause mass media campaigns to succeed or fail. Research into these processes has so far been largely atheoretical, however, and has examined responses to messages in the aggregate. Thus, little is known about the ways these patterns of neural activity might be affected by specific features of the persuasive messages that induce them.

Dissertation overview

This dissertation seeks to improve the understanding of message-driven persuasion by uncovering neural mechanisms that underlie responses to certain theory-relevant message features. Specifically, the project tests the neural correlates of message features across three independent studies.

Chapter two seeks to build on past findings that neural correlates of subjective value responded to the gain/loss framing of persuasive messages (Vezich et al., 2017). Responses to gain/loss framing in various contexts are described by the framing postulate of prospect theory (Kahneman & Taversky, 1973; Tversky & Kahnemann, 1989), which explains the ways the

people make decisions under varying conditions of perceived risk. While the effects of gain/loss framing on message outcomes have been tested extensively in health communication (Rothman et al., 2020) and prospect theory has been used to explain neural activity in neuroeconomics research (Trepel, Fox & Poldrack, 2005), this chapter is among the first studies to explore the ways gain/loss framing of health communication messages affects the activity they elicit in the brain.

Chapter three explores the neural correlates of reactance processes (Brehm, 1966) in response to a set of real-world PSAs that use either humor or fear appeals. Past research suggests that humor appeal messages elicit less reactance than fear appeal messages (Zhao, Roditis, & Alexander, 2019), but little is known about the ways that reactance processes (i.e. freedom threat, anger, and negative cognitions) manifest themselves in the brain. By comparing responses to the same set of messages across independent fMRI and survey samples, this study tests whether messages that elicited more activity in brain regions associated with reactance processes also elicited stronger self-reported experiences of those processes. This study also tests the effects of humor/fear appeals on both neural correlates of and self-reported levels of reactance processes.

Chapter four tests whether and how the additions of warning labels to social media posts promoting cannabis might reduce people's willingness to share those posts online. This study tests the effects of warning labels in an online survey and examines activity in brain regions that are important to a neural model of information virality (Scholz et al., 2020), making it the first exploration into the effects of message features on this model. Importantly, online sharing of cannabis marketing materials has been described as a critical public health risk (Moreno, 2022), so this study tests theorized persuasive mechanisms in a context that could directly inform public

health policy.

Finally, the findings from each of these studies are briefly summarized in chapter five and discussed in the context of the larger theoretical issues described in this introduction.

CHAPTER 2: GAIN/LOSS FRAMING MODERATES THE vmPFC'S RESPONSE TO PERSUASIVE MESSAGES WHEN BEHAVIORS HAVE PERSONAL OUTCOMES

Introduction

More than a decade of research has shown that activation of the ventral portion of the medial prefrontal cortex (vmPFC) in response to persuasive messages predicts message-consistent behaviors and other desired outcomes, accounting for variance in behavior that is unique from self-report measures of message effectiveness (Falk et al., 2010, 2011, 2012, 2015a, 2015b; Cooper et al., 2015; Vezich et al., 2017; Pandey et al., 2021). This relationship has been established at two different levels. At the participant level, individuals who show more vmPFC activation during message exposure are more likely to show desired behavior changes after message exposure (Falk et al., 2010). At the message level, messages that elicit more vmPFC activation in small scanner samples tend to be more successful at the population level (Falk et al., 2012). Researchers have suggested that increased vmPFC activation underlies perceptions of message value in relationship to one's goals and motivations (Falk & Scholz, 2018), but little is known about the ways this process might be affected by features of the messages themselves (Falk et al., 2015). This has made it difficult for scholars to connect neuroimaging research with existing theories of persuasion (Vezich et al., 2016) and develop evidence-based recommendations for message design.

To bridge this theoretical gap, the current study sought to build upon recent research into effects of a common message-level feature: gain/loss framing. Specifically, the current study follows a line of inquiry opened by Vezich et al. (2017), who examined vmPFC activation in response to messages that promoted sunscreen use. These researchers found that gain-framed messages (i.e., messages that emphasized potential benefits of using sunscreen) elicited more vmPFC activation than loss-framed messages (i.e., warnings about the risks of not using

sunscreen), and that the strength of this gain/loss framing effect on vmPFC activation predicted actual sunscreen use even when controlling for behavioral intentions (Vezich et al., 2017). Importantly, this effect was only observed in responses to a "why"-framed subset of the messages (i.e., messages explaining why one should wear sunscreen). When researchers attempted to replicate this effect using functional near infrared spectroscopy (fNIRS), they observed a positive association between mPFC activation and sunscreen use but did not find evidence that mPFC activation was sensitive to the gain/loss framing of messages (Burns et al., 2018).

Given the prominence of gain/loss message framing in health communications research (Guenther, Gaertner & Zeitz, 2021), inconsistent findings within the original Vezich et al. (2017) study between "how" and "why" messages, and the observation of null gain/loss framing effects in the attempted replication (Burns et al., 2018), there is a need to furer examine how gain- and loss-framed messages differ in their effects on the vmPFC, a key region associated with effective messaging and behavior change (Falk & Scholz, 2018). Further, it remains unknown whether the gain/loss framing effect can be observed outside the context of messages about sunscreen use. Therefore, the current study aimed to examine whether vmPFC activity differed in response to gain- versus loss-framed messages that varied across two dimensions: the temporal orientation of described outcomes (present-oriented vs. future-oriented), and the nature of those outcomes (personal vs. prosocial).

Persuasion and vmPFC activation

Activation of various subregions of the mPFC in response to persuasive messages has been associated with desired outcomes across a variety of messaging contexts (for a review, see; Cacioppo, Cacioppo, & Petty, 2018). For example, research examining vmPFC activity during

exposure to messages that presented expert information about the benefits of sunscreen use compared to rest found that participants who displayed increased vmPFC activation when processing the appeals were more likely to increase sunscreen use in the week after the scan (Falk et al., 2010). In addition, increased mPFC activity among smokers exposed to tailored cessation messages (compared to untailored messages) predicted abstention from cigarettes a full four months after the scan session (Chua et al., 2011). Similar effects have been observed among participants exposed to messages encouraging physical activity (Cooper et al., 2017) and to pictorial warning labels designed for cigarette packages (Riddle et al., 2016). In all cases, vmPFC activation predicted variance in outcomes of interest over and above the variance explained by self-report measures.

Researchers have also found that vmPFC activation predicts desired outcomes at the message level. For example, when Falk et al. (2012) examined responses to three sets of anti-smoking PSAs, they found that a rank ordering of the campaign messages by average vmPFC activation aligned with a rank ordering of the call volume to a tobacco quitline in the months after each campaign was broadcast (driving smokers to call the quitline was the intent of all campaigns). Interestingly, this rank ordering was different from the order both participants and experts proposed when they were asked to rank the campaigns by likelihood of success—the message that participants and experts ranked as the least likely to succeed ultimately elicited the highest average level of vmPFC activation and the greatest number of calls to the quitline. Similarly, research examining the relationship between smoking cessation materials and email click-through rates found that materials that elicited increased vmPFC activation in an fMRI sample also had higher click-through rates when distributed at the population level through an email newsletter (Falk et al., 2015b). Once again, participants' PME ratings were not effective

predictors of message success.

In sum, activity within the vmPFC in response to persuasive messages has been shown to predict both individuals' tendencies to engage in message-consistent behaviors and the tendency of messages to elicit desired outcomes in real-world settings (Falk et al., 2010; Falk et al., 2012). Further, the tendency of vmPFC activity to outperform self-report as a predictor suggests that this measure might capture some message effects that occur outside of participants' conscious awareness.

Because no brain region is exclusively associated with any one cognitive process, it is difficult to infer an individual's cognitive experience from activation alone (Poldrack, 2006). Having witnessed its activation in several similar structured contexts, however, Falk and Scholz (2018) have proposed that vmPFC activation elicited by persuasive messages indicates a process of *subjective valuation*, which they assert is strongly associated with perceptions of self-relevance when audiences process persuasive messages. In other words, they proposed that vmPFC activity elicited by persuasive messages indicates that audience members perceive the value of the message as being relevant to them personally. Similarly, Vezich et al. (2017) suggested this activity might signal a process of *self-value integration*, in which individuals identify the target behavior as valuable and choose to incorporate it into their self-concept. Assertions about the nature of this process remain hypothetical, but a growing body of work suggests that vmPFC activity is a reliable predictor of persuasive message success.

Prospect theory and gain/loss messages

First proposed in the field of behavioral economics, the framing postulate of prospect theory (Kahneman & Taversky, 1973; Tversky & Kahnemann, 1989) asserts that people tend to be risk-averse when considering the benefits of an action but are willing to tolerate risk when

motivated to avoid a negative outcome. In the context of health promotion, existing evidence suggests that loss-framed messages are more effective at promoting detection behaviors that might involve undesirable outcomes, such as cancer screening and testing for sexually transmitted infection (Rothman et al., 2020; Rothman & Salovey, 1997). On the other hand, messages promoting relatively low-risk behaviors such as increases in physical activity tend to be more effective when they emphasize potential gains (Rothman et al., 2020; Rothman & Salovey, 1997). Given that sunscreen use is a relatively low-risk behavior, Vezich et al.'s (2017) finding that gain-framed messages on this topic tended to elicit increased vmPFC activation is consistent with prospect theory's framing postulate. Still, it remains to be seen whether this effect might also be observed in different combinations of message features.

Message features

Whenever researchers observe an effect within an experimental sample, they must consider sources of heterogeneity among participants before generalizing that effect to a larger (and likely more diverse) population of people. Similarly, researchers hoping to identify generalizable message-level effects must account for the reality that persuasive messages can differ in many ways. At present, most research on neural processes associated with persuasion combines multiple message features, which could explain why average-level message effects have been identifiable, but have included positive, negative, and null effects (O'Keefe & Hoeken, 2021). To explore how gain/loss framing effects might generalize to a more diverse set of relevant messages, we varied our sample of messages across two additional dimensions: the temporal orientation of the gain or loss outcome and the nature of that outcome.

Temporal orientation of the outcome (present vs. future)

When describing the likely outcomes of adopting or failing to adopt a certain behavior,

message designers often must choose between immediate (present-oriented) or distal (future-oriented) outcomes. For example, sunscreen use can be associated with either present-oriented gains like social approval or future-oriented gains like younger-looking skin later in life. Similarly, failure to use sunscreen can be associated with both present-oriented losses like sunburn and future-oriented losses like an increased risk of skin cancer.

Past research has shown that changes in temporal orientation can impact the effectiveness of messages that advocate for both low-risk (Strathman et al., 1994) and detection (Kreuter et al., 2005) behaviors, and that the direction of these effects depends on traits of a message's audience. Therefore, the present study tested for gain/loss framing effects in messages that oriented outcomes in the present and in messages that oriented outcomes in the future.

Nature of the outcome (personal vs. prosocial)

Another important difference between persuasive messages is the nature of the behavior advocated for. Messages often focus on behaviors that impact individuals personally—for example, the outcomes of improving diet or failing to apply sunscreen will be felt primarily by the person who does or does not engage in those behaviors—but this is not always the case. Some messages advocate instead for prosocial behaviors, for which risks are incurred by the individual but benefits are experienced by another or by the community at large. For example, Helme and colleagues (2020) showed that Appalachian residents perceived preventing accidental and intentional misuse of opioids by others in their household as a benefit to disposing of unused opioid medications instead of keeping them in the home. Additionally, potential organ donors perceive the opportunity to save a life as a benefit of donation (Quick et al., 2014; Siegel et al., 2010; Williamson et al., 2017).

The study of behavior change has increasingly shown interest in the health and

environmental domains (Liang et al., 2018) which also offer prosocial benefits (e.g., organ donation) or a mix of personal and prosocial benefits (e.g., vaccination). However, the field has not engaged in much theorizing and direct comparison of personal and prosocial benefits of behaviors. Work in environmental psychology gives some insight into the unique challenges associated with prosocial behavior, including difficulties thinking beyond the self and drawing a connection between one's own behavior choices and positive social impact (Gifford, 2011; Liang et al., 2018). Therefore, the context of relevant outcomes (i.e. whether they were framed as applying to the individual or to some external agent) is a dimension along which messages often vary. In the present study, this was explored by testing messages that advocated for two types of behaviors: physical exercise (personal) and pro-environmental behaviors (prosocial).

The current study

The current study seeks to expand on the finding of Vezich et al. (2017) that gain-framed messages tend to elicit increased activation in the vmPFC compared to loss-framed messages encouraging adoption of a low-risk behavior. We sought to test whether this effect could be generalized to a set of messages that varied across some common dimensions (present-/future-oriented, personal-/prosocial-focused). In addition, given that self-report ratings of PME have varied in how well they correspond to neural activity during message encoding (Falk et al., 2012, Falk et al., 2015), we tested both the effects of gain/loss framing on PME and the associations between PME and vmPFC activity. We hypothesized that gain-framed messages would elicit increased activation of the relevant vmPFC subregion and higher levels of PME compared to loss-framed messages across all possible message conditions.

Methods

Participants

Forty-five participants were recruited through an online job board from a large midwestern research university, aged 18-34 years old (M = 20.53 , SD = 2.65; N_{Female} = 30). All participants were right-handed, did not suffer from claustrophobia, had normal (or corrected to normal) vision, were not taking any psychoactive medications and did not have any history of psychiatric or neurological disorders. Data from one participant was excluded from this analysis due to errors during data collection, leaving a final sample of 44 participants.

Given concerns about low statistical power in neuroimaging studies (Yarkoni, 2009), we used the software G*Power to test whether our sample size was sufficient to detect the effects observed in past research at statistical power at .90, α = .05 using one-sample t-tests. To observe an estimated effect size d = 1.22, as calculated using the six R-squared values in Burns et al. (2018), the recommended total sample size was eight participants. To detect an estimated average effect size d = 0.62, which we calculated using the nine Pearson's r values in Vezich et al. (2017), the recommended total sample size was 29 participants. Thus, our sample of N = 44 was deemed sufficient to detect similar effects.

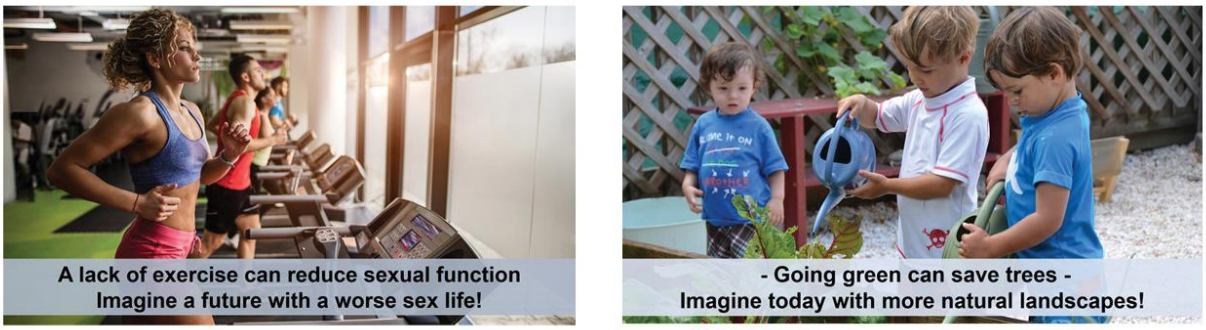
Procedure

After obtaining consent, participants completed a series of pre-scan questionnaires (unrelated to this study), fMRI safety screening, and were trained on fMRI at a research facility on the campus of a large midwestern university. Participants then underwent a one-hour scan session that included a nine-minute persuasive message task. All components of this protocol were approved by the relevant university IRB.

Persuasive message tasks

While undergoing fMRI, participants viewed a series of 44 randomly-ordered persuasive messages in a rapid event-related design. Messages were constructed for this experiment and varied across three fully crossed conditions: each used either a gain or loss frame (gain/loss conditions), described the outcomes of adopting or failing to adopt the behavior in either the present or the future (present-oriented/future-oriented conditions), and advocated either for an increase in physical activity or adoption of pro-environmental behaviors (personal/prosocial conditions). Two examples are presented in Figure 2.1.

Figure 2.1: Examples of persuasive message stimuli



Note: *Left: A loss-framed, future-oriented message encouraging physical activity. Right: A gain-framed, present-oriented message encouraging pro-environmental behavior.*

Participants viewed each message for a period of five seconds, then were asked "How effective is this message?" to rate their perceptions of the message's effectiveness (PME) on a scale of 1 to 4 using a four-button Current Designs response pad held in the participant's right hand. Participants were given three seconds to enter a rating, then viewed a fixation cross for a period of 1.5 seconds before presentation of the next message.

Framing manipulation check

Manipulation of the gain/loss framing of stimuli in this study was performed by making

changes to the text that accompanied each image, but each textual message was always paired with the same image. To ensure that any effects of gain/loss framing were driven by manipulation of the text instead of some feature of the images used, we conducted a follow-up online survey to a demographically similar sample of *N*=238 participants recruited via Qualtrics Panels. For the online study we altered the 22 pro-activity messages by reversing the frame associated with each image. Thus, each image that was associated with a gain-framed textual message in the original study was associated with a loss-framed textual message in the follow up survey, and vice versa. Participants were prompted to provide a PME rating to the prompt: "How effective is this message?" Response options ranged from 1(Not at all effective) to 5(Extremely effective), similar to the fMRI study.

fMRI data acquisition and preprocessing

Structural and functional brain imaging was conducted using a 3 Tesla GE Discovery MR750 scanner. Head motion was minimized using foam padding on the head coil. Two functional runs were recorded (TR = 800ms, TE = 20, flip angle = 60°, matrix size = 96x96, 54 axial slices, 3mm thick; voxel size = 3.0x3.0x3.0), and a motion-corrected T1-weighted MPnRAGE acquisition with 1.0 mm isotropic spatial resolution was used as an anatomical underlay (Kecskemeti et al., 2018). Image preprocessing and analysis were performed using the afni_proc.py program within the Analysis for Functional Neuroimaging (AFNI) software package (Cox, 1996). Functional and anatomical runs were warped to align with the Talairach-Tournoux N27 template brain and smoothed with a 4mm Gaussian kernel. To ensure only steady-state images were used in our analysis, we discarded the first 7 TRs (5600 ms).

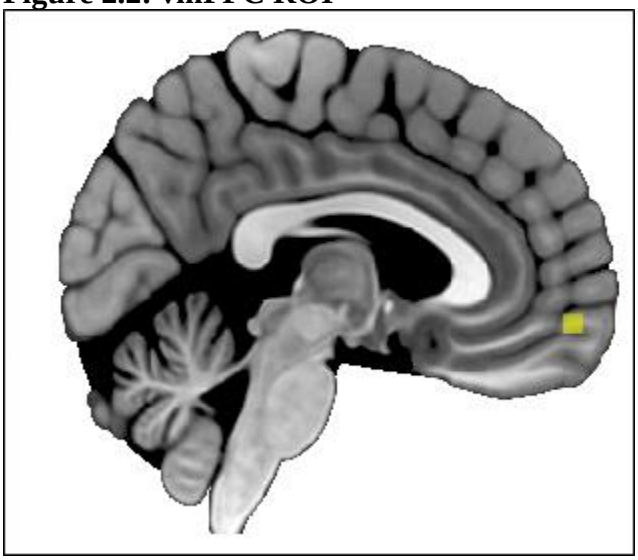
To assess the effects of gain/loss framing in the full dataset and in response to four relevant subsets, we estimated first-level regression models for the gain and loss conditions

across the full set of messages and for message subsets of interest, including pro-activity, pro-environmental, present-oriented, and future-oriented messages. Each model included instances of both gain- and loss-framed messages as regressors, as well as a regressor for the parametric modulation of neural signal by PME. Random effects, motion, and nuisance regressors were also modeled using AFNI's 3dDeconvolve command.

Region of interest analyses

We were interested in a single *a priori* region of interest: a portion of the vmPFC overlapping the ROI used by Vezich et al. (2017). Activation of that region in response to persuasive messages has been used to predict behavior in a variety of contexts (Falk et al., 2011, 2012; Cooper et al., 2015, Falk et al., 2015). Like Vezich and colleagues, we focused specifically on the cluster that was found to predict behavior in Falk et al., 2010. Specifically, we constructed a 10mm, 33-voxel sphere around the coordinate 0,60,-9 using the Montreal Neurological Institute template brain (Figure 2). From this ROI, we exported four different measures for our analyses: gain/loss framing effects, present/future orientation effects, personal/prosocial outcome effects, and parametric modulation by PME.

Figure 2.2: vmPFC ROI



Note: *The vmPFC ROI was constructed using a 10mm, 33-voxel sphere centered around the coordinates x = 0, y = 60, z = -9 using the Montreal Neurological Institute template brain.*

We estimated gain/loss framing effects for each of our four message design categories: present personal, future personal, present prosocial, and future prosocial. To do this, we performed a General Linear Test comparing signal during exposure to gain-framed messages with activation during exposure to loss-framed messages for each category, then averaged coefficients for all voxels within each participant's vmPFC ROI. We estimated present/future and pro-activity/pro-environmental effects in a similar manner, comparing present-oriented messages with future-oriented messages and personal with prosocial messages, respectively. Finally, we tested for associations of ROI activity and PME ratings using a parametric modulation regressor.

Statistical tests

Tests of gain/loss framing on vmPFC activity

First, we tested whether the vmPFC was sensitive to the gain/loss framing messages by conducting a series of one-sample t-tests on our measure of the gain/loss framing effect within

our ROI. We performed four independent tests—one for each of our message conditions: present personal messages, future personal messages, present prosocial messages, and future prosocial messages.

Second, we tested whether the effects of gain/loss framing on vmPFC activity were different for different combinations of message features by testing the fit of a series of linear mixed effects regression models. Each model used our measure of gain/loss framing effects in the vmPFC as the outcome and included dummy-coded regressors for three of the four possible message conditions. To account for the nested nature of the data, each model also included a by-participant random intercept and a by-participant slope for each dummy-coded regressor. Three models were tested in all, allowing for pairwise comparisons of all message conditions.

Third, we tested whether our other manipulated message features had any effects on vmPFC by conducting one-sample t-tests on our measures of present/future orientation effects and personal/prosocial effects.

Tests of effects of message conditions on PME

To test the effects of our message conditions on participants' PME ratings, we fit a linear mixed effects model in which PME ratings were the outcome and our gain/loss, present/future ,and personal/prosocial conditions and their interactions were entered as fixed effects. To account for the nested nature of our data, we also included by-participant and by-stimulus random intercepts and random slopes for all fixed effects (models that included by-participant or by-stimulus random slopes for interaction effects failed to converge).

Test of relationship between vmPFC activation and PME

To test whether vmPFC activation during message encoding was associated with

participants' subsequent rating of the messages, we conducted a one-sample t-test on the ROI-wide coefficients for our parametric modulation regressor.

Test of framing manipulation check

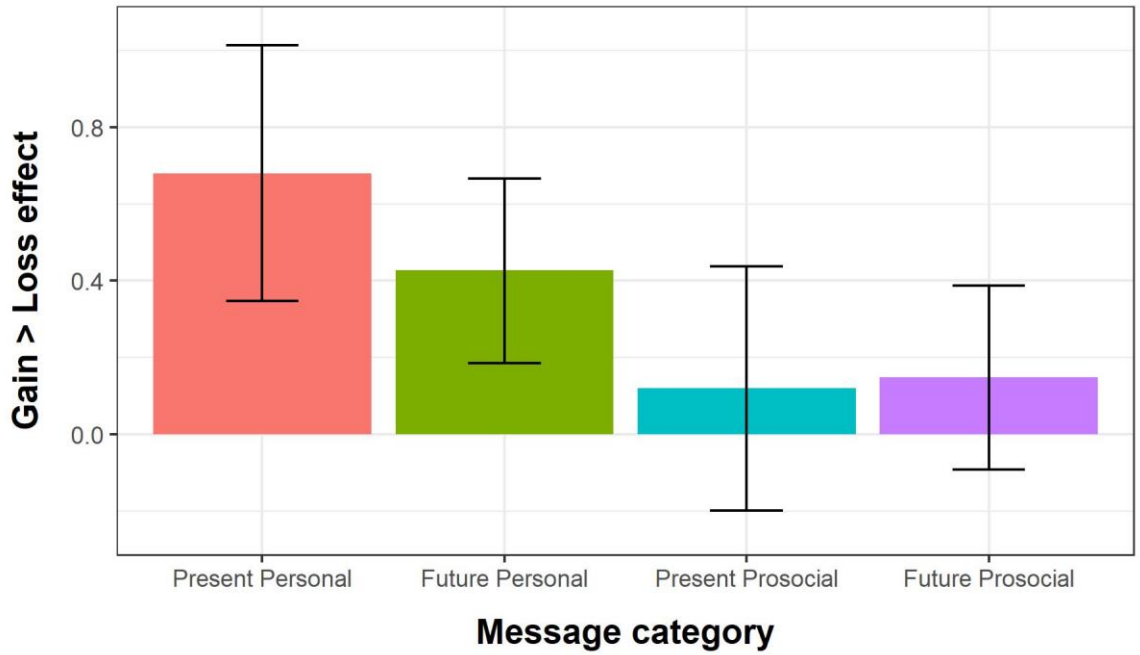
To confirm that gain/loss framing effects were not driven by the images used in our stimuli, we conducted a test on our independent survey sample in which the images used were reversed. We tested the fit of a linear mixed-effects model with fixed effects for gain/loss framing and present/future orientation, by participant random slopes for each fixed effect, and a by-participant random intercept.

Results

Effects of gain/loss framing on vmPFC activity

First, we tested whether the gain/loss framing of messages impacted the level of activation elicited within the vmPFC for each of our message categories. Results indicated that neural activity within the vmPFC was significantly higher in response to gain-framed messages than loss-framed messages for two of our four message categories: *present personal* messages ($t(43) = 4.062$, $p < 0.001$, $p_{adjust} < 0.001$ 95% CI[0.343, 1.018]) and *future personal* messages ($t(43) = 3.526$, $p = 0.001$, $p_{adjust} = 0.005$, 95% CI [0.183,0.671]). We did not find evidence that gain/loss framing had any effect on vmPFC activity during encoding of either *present prosocial* ($t(43) = 0.747$, $p = 0.357$, $p_{adjust} = 0.714$, 95% CI[-0.203, 0.442]) or *future prosocial* messages ($t(43) = 1.208$, $p = 0.459$, $p_{adjust} = 0.714$, 95% CI[-0.013, 0.052]). All p-values were adjusted for multiple comparisons using the Holm-Bonferroni method. Results of these tests are illustrated in Figure 2.3 below.

Figure 2.3: Effects of gain/loss framing across message conditions



Note: *X-axis describes message category tested, Y-axis describes average coefficient in the vmPFC from the contrast gain>loss. Error bars represent 95% confidence intervals.*

Second, we compared the strength of gain/loss framing effects across our message categories by testing a series of linear mixed effects models. These tests allowed us to compare each possible pair of categories. We found that the gain/loss framing effect in responses to *present personal* messages was significantly stronger than for *future prosocial* messages (β = 0.532, $F(1,48.818) = 7.963$, $p = 0.007$, $p_{adjust} = 0.042$), and was stronger than *present prosocial* messages, though this effect did not remain significant after our multiple comparison correction (β = 0.561, $F(1,52.367) = 6.342$, $p = 0.015$, $p_{adjust} = 0.075$). We did not find evidence that the strength of the gain/loss framing effect differed across any other pair of message conditions. Complete results are presented in Table 1 below.

Finally, we tested whether activity in the vmPFC was affected by either of our other message manipulations: present/future orientation or personal/prosocial outcomes. We did not observe significant differences in vmPFC activity when comparing present-oriented message

exposures to future-oriented message exposures ($t(43) = -0.538$, $p = 0.999$, 95% CI [-0.024, 0.014]), and the difference we observed between messages with personal outcomes vs. those with prosocial outcomes did not survive our multiple comparison correction ($t(43) = 2.28$, $p = 0.027$, $p_{adjust} = 0.110$, 95% CI [0.003, 0.050]).

Effects of gain/loss framing and other message features on PME

Next, we tested whether gain/loss framing and message features affected participants' PME ratings. Tests of our mixed-effects model suggests that gain/loss framing did exert a significant effect on PME ratings, such that participants tended to offer higher PME scores for gain-framed messages than for loss-framed messages ($\beta = 0.616$, $F(1,60.099) = 34.382$, $p < 0.001$).

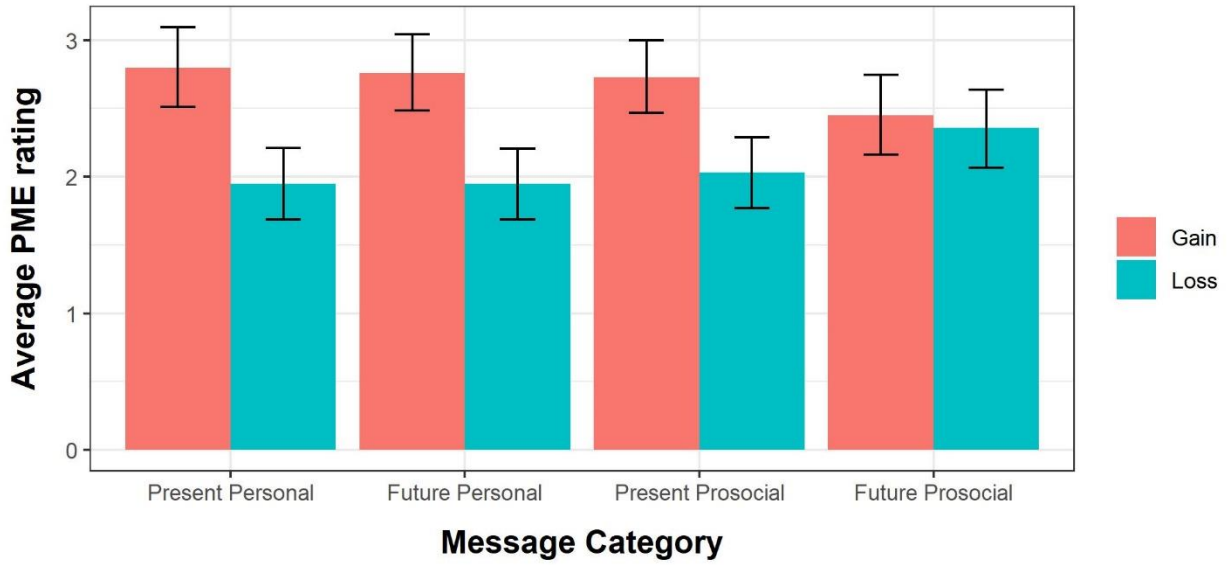
We also found that gain/loss framing interacted with our present/future condition ($\beta = 0.309$, $F(1,1762) = 4.535$, $p = 0.041$), such that the difference in PME ratings between gain-framed and loss-framed messages was larger for messages with present-oriented outcomes compared to future-oriented outcomes. Gain/loss framing also interacted with our personal/prosocial condition ($\beta = 0.432$, $F(1,30.244) = 8.756$, $p = 0.006$) such that the difference in PME ratings between gain-framed and loss-framed messages was larger for personal messages than prosocial messages Compete results are presented in Table 2.1 and illustrated in Figure 2.4.

Table 2.1: Effects of message categories on perceived message effectiveness

	β	*F*	*Df*	*Df.resid*	*p*
Gain/Loss(GL)	0.614	34.382	1	60.099	<0.001
Present-oriented/Future-oriented (PF)	0.005	0.004	1	33.580	0.948
Personal/Prosocial(PP)	0.026	0.094	1	43.766	0.760
*GL*PF*	0.312	4.535	1	30.243	0.041
*GL*PP*	0.434	8.755	1	30.244	0.006
*PP*PF*	-0.069	0.223	1	30.253	0.640
*GL*PP*PF*	-0.595	4.128	1	30.251	0.051

Note: *Results of a linear mixed-effects model testing the effects of message features on perceived message effectiveness (PME). Significant results indicate that PME ratings differed significantly across message subsets.*

Figure 2.4: Mean PME ratings for gain- and loss-framed messages across message categories



Note: *X-axis describes message category, Y-axis describes mean PME ratings by framing condition. Error bars represent 95% confidence intervals.*

Association between PME and vmPFC activity

We tested the coefficients of our parametric modulation regressor to determine whether vmPFC activity during message encoding was associated with subsequent PME ratings. Results

indicated that activation within the vmPFC was positively associated with PME ratings ($t(43) = 2.592$, $p = 0.013$, 95% CI [0.005, 0.042]). In other words, participants showed more vmPFC activity while encoding messages they subsequently rated as effective.

Framing manipulation check

Finally, we checked whether the effect of gain/loss framing was driven by the images paired with the manipulated text in our stimuli by testing the fit of a linear mixed effects model against an independent sample in which the images were reversed. Results revealed an effect of gain/loss framing in the same direction as the fMRI findings (B=0.437, $F(1,237.02)$=58.9, p<0.001), suggesting that differences in PME and vmPFC activity were driven by message frames, rather than attributes of the images.

Discussion

To better understand the role of message features in the neural processing of persuasion, this study tested whether vmPFC activity during message encoding varied depending on whether messages framed outcomes as gains or losses. Past work found that certain gain-framed messages elicited more vmPFC activity than similar loss-framed messages (Vezich et al., 2017), but it remained unclear whether this effect would generalize to messages with other combinations of message features. Our results suggest this effect can be observed in response to a diversity of messages but may depend on whether those messages address outcomes relevant to individual persons or to society at large. We also found that gain/loss framing influenced participants' ratings of perceived message effectiveness, and that these ratings were positively associated with activity in the vmPFC. Taken together, our results align with theorizing that vmPFC activity during persuasive processing indexes a process of self-value integration, but they may suggest that this process only involves or depends on some assessment of personal risk.

Gain/loss framing affects vmPFC activity in messages about personal outcomes

Using an ROI modeled after past work (Vezich et al., 2017) that encompasses a region associated with successful persuasion (Falk et al., 2010; 2012), we observed that certain gain-framed messages elicited more vmPFC activity than similar loss-framed messages. This effect emerged regardless of whether messages framed outcomes in the present or in the future, but we did not observe this effect in response to messages advocating behaviors with prosocial outcomes. Further, a comparison of gain/loss framing effects across message conditions found stronger effects in one personal outcome condition (present personal) than in both prosocial outcome conditions. These results clarify the findings of Vezich et al. (2017), suggesting the effect reported in their paper may be generalizable to a diverse population of messages despite a previous failure to replicate (Burns et al., 2018). Further, they suggest that established theories of persuasion like prospect theory (Kahneman & Tversky, 1973; Tversky & Kahneman, 1989) may offer useful frameworks for explaining the relationships between the features of persuasive messages and the neural processes they evoke.

Our finding that gain/loss framing effects could only be observed in response to messages with personal outcomes has two possible explanations. First, the prosocial outcome messages may have contained some design features that were not intentionally manipulated, but nonetheless interfered with the gain/loss framing effect. For example, it is possible that prosocial messages were generally of lower quality than personal messages. Comparison of message PME scores suggests this is unlikely: overall PME ratings did not differ significantly across the present-oriented/future-oriented or personal/prosocial message conditions. Although PME and vmPFC activation are distinct measures, the absence of any differences in PME ratings across message conditions suggests differences in activation are likely not due to major differences in

message quality.

Another explanation is that the vmPFC's sensitivity to gain/loss framing of messages depends on whether those messages address personal vs. prosocial outcomes. The messages used in this study shared important similarities with those tested by Vezich et al. (2017), who also observed gain/loss framing effects during encoding of some messages but not others ("why"-framed messages but not "how"-framed messages). All messages used in this study were "why"-framed and advocated for behaviors that, like sunscreen use, are considered low risk (increasing physical activity and "going green"). These similarities are appropriate given the tenets of prospect theory, which proposes that people are more attentive to gains than losses when they do not consider themselves to be at risk but more attentive to losses when they do feel at risk. Thus, the current study manipulated two message dimensions that were not expected to affect risk perceptions: the temporal orientation of the outcomes and the objects of those outcomes. Results suggest that gain/loss framing affects vmPFC activation regardless of whether a message focuses on the near-term or long-term outcomes, but it does not appear to have the same effect when messages describe gains or losses that do not apply to their audience directly.

This latter explanation aligns with theorizing that the vmPFC's role in persuasion processing is to incorporate perceptions of message value with considerations about the self. Vezich et al. (2017) proposed that gain-framed messages elicited more activity because they made the personal value of a behavior more explicit, making participants more likely to reflect positively on the prospect of incorporating the behavior into their self-concept. We suggest that our observations captured a similar process, but that the absence of effects in prosocial conditions might indicate a slightly different form of self-related processing. Given that prospect theory suggests framing effects depend on perceptions of relative risks and rewards, we propose

that our personal/prosocial manipulation impacted some component of these risk perceptions. The absence of differences in vmPFC activity across present/prosocial conditions suggests that this activity does not index only immediate self-relevance, but such relevance might be an important antecedent to value judgements about message content. In this case, the "self" processes included in self-value integration might involve some consideration not just of self-concept but of the outcomes themselves. For example, participants might have formed mental simulations of messages' outcomes and shown vmPFC activity only with those simulations involved themselves personally.

Gain/loss framing affects perceived message effectiveness

Gain/loss framing also exerted an effect on PME ratings such that participants tended to rate gain-framed messages as more effective than loss-framed messages. We did not observe any direct effects on PME by other message features, but these features interacted with gain/loss framing so that its effects on PME were stronger when messages presented present-oriented and personal outcomes. Notably, this aligns with our comparisons of gain/loss framing effects on vmPFC activity across message conditions, which found that framing effects were significantly stronger during encoding of present-oriented personal messages than future-oriented prosocial messages. Thus, we found that gain/loss framing was a uniquely important factor in perceptions about a message's effectiveness, and that the way these perceptions varied across our message conditions aligned in some ways with variance in the gain/loss framing effects on vmPFC activity.

The link between the vmPFC and PME is further evidenced by the test of our parametric modulation regressor, which found that participants exerted more vmPFC activation when encoding messages they subsequently rated as more effective. This aligns with other work that

has found activity in this region can predict not only value judgments but also assessments of personal significance. For example, when D'Argembeau et al. (2012) asked participants undergoing fMRI to reflect on their personal traits, they observed greater vmPFC activation when participants reflected on traits they considered personally or emotionally significant. Similarly, Lin et al. (2016) tested a parametric modulation regressor like our own to find vmPFC activity during recall of autobiographical memories tracked with the degree to which those memories involved things that participants valued or liked. Activity in this region has also been shown to track with value judgments of less personal stimuli, including monetary rewards (Shapiro & Grafton, 2020), making it widely considered a domain-general index of value (Bartra et al., 2013). Thus, our finding that vmPFC activity during message encoding tracks with subsequent PME ratings is consistent with past research suggesting this region is involved in value judgments.

Suggestions for future research

These findings suggest several promising opportunities for future research. First, this work could investigate the possibility that vmPFC activity during persuasion involves or depends on some assessment of self-relevant risks. This is one potential explanation for our finding that the vmPFC was only sensitive to the gain/loss framing of messages that addressed personal outcomes, and it aligns with some components of prospect theory. Perceptions of risk are an important consideration of prospect theory, which posits that people are more attentive to gains when they perceive risks as low and more attentive to losses when they perceive risks as high. Thus, future work could both test our suggestion that risk assessments are important to self-value integration processes and further test the explanatory power of prospect theory by experimentally manipulating the degree to which messages are likely to induce perceptions of risk. Second,

future work could incorporate independent measures of message success, such as message-consistent behavior or population-level message performance. This could help to further establish that gain-framed messages are more effective and that differences in vmPFC activation across gain-vs-loss-framed messages suggests differences in the real-world effectiveness of those messages. Finally, testing of messages advocating for a variety of prosocial behaviors and future-oriented behaviors would provide context for our findings that differences in the gain/loss framing of prosocial messages do not tend to elicit the expected differences in vmPFC activation for messages of these types.

Conclusions

The current study found that gain-framed messages advocating for a low-risk behavior (physical activity) and present harms/benefits (present orientation) elicited increased vmPFC activity compared to loss-framed messages. Participants also tended to rate these messages as more effective. These results replicate past work (Vezich et al., 2017), contributing to nascent efforts to bridge neuroscientific research with established theories of persuasion. This suggests a promising future for research of this kind, which may soon allow neuroscientists to offer empirically-supported recommendations for message design.

CHAPTER 3 – TESTING FOR NEURAL CORRELATES OF REACTANCE IN RESPONSE TO ANTI-DUIC PSAS USING FEAR AND HUMOR APPEALS

Introduction

First legalized by California for medical use in 1996, cannabis is now legal in some form in 39 U.S. states and is fully legal for adults in 23 states and in the District of Columbia (Marijuana Policy Project, 2023). Adults in the United States have become increasingly supportive of cannabis legalization over the last 30 years (Chiu et al., 2022), and have come to view cannabis products as having fewer health risks (Carliner et al., 2017). Cannabis products remain illegal under federal law, but their use is increasingly becoming a normal and accepted part of American life. Thus, state and local governments are often tasked with establishing, enforcing, and communicating norms and guidelines around behaviors that have long been practiced outside legal bounds.

One behavior that has attracted the attention of public health agencies is driving under the influence of cannabis, or DUIC. Cannabis use is known to impair driving abilities, particularly among infrequent smokers (Hartman & Heustis, 2013), and cannabis impairment likely causes thousands of motor vehicle deaths each year (WHO, 2016a). Incidents of DUIC and cannabis-related motor vehicle incidents have increased recently in both U.S. (Myers et al., 2023), and research suggests that legalization of cannabis is associated with increases in the rates of fatal motor vehicle accidents involving cannabis (Windle et al., 2021). Thus, DUIC is increasingly considered a serious threat to public health and safety (Ramaekers, 2018; WHO, 2016a).

Many states have sought to prevent increases in DUIC rates through mass media campaigns—television public service announcement (PSA) spots on the topic have been produced by government agencies in California, New Mexico, and Colorado, and by the National

Highway Safety Traffic Administration. Similar campaigns have proven effective in reducing rates of driving under the influence of alcohol (Elder et al., 2004), but little is known about the effects of campaigns advocating against DUIC. A recent scoping review found that mass media campaigns about DUIC were largely unexamined in the existing literature, with only three being described in peer-reviewed studies and only one examined through a theoretical lens (Colonna, 2022). Best practices for the design of mass media campaigns for traffic safety suggest that designers consult existing research on past campaigns and consider theories of behavior change communication while developing their campaigns (Noar et al., 2006; Delhomme et al., 2009). Thus, it is important to develop a body of theory-driven research into the effects of mass media campaigns about DUIC.

One theoretical perspective that may be useful in this context is psychological reactance theory (PRT; Brehm, 1966), which explains a mechanism that sometimes causes audiences to reject persuasive messages by presenting a threat to their perceived freedom to choose. Though it has yet to be examined in the context of anti-DUIC messages, psychological reactance has been suggested as an explanation for the failure of past mass media campaigns about cannabis (Hornik et al., 2008) and has been shown to impact responses to traffic safety messages (Ward et al., 2021). However, the underlying cognitive processes associated with PRT have largely been inferred from self-report and lack objective evidence (Rains, 2013). Thus, the present study sought to extend our theoretical understanding of both anti-DUIC messaging and PRT by examining the relationship between neural activity during anti-DUIC media messages and PRT.

This study has four aims: (1) to test whether an established model of reactance is associated with responses to a set of anti-DUIC messages, (2) to test whether anti-DUIC messages are associated with synchronized neural activity across viewers, (3) to examine

whether synchronized neural activity across viewers during anti-DUIC messages is associated with a model of reactance, and (4) to test whether reactance processes are affected by the type of appeal used by an anti-DUIC message. The theoretical and real-world significance of these questions are established in the literature review below. In that review, I first introduce PRT and the intertwined model of reactance, which constitutes the theoretical framework of this study. Second, I explain the real-world significance of reactance in the context of anti-cannabis and anti-DUIC messaging. Third, I describe a style of neuroimaging analysis that allows researchers to observe the synchronized neural responses produced by messages and explain how this method has previously been used to understand message effects. Fourth, I present a detailed overview of three key PRT processes–freedom threat perceptions, anger, and negative cognitions–and offer hypothesized neural correlates for each. Fifth, I explain the differences between fear and humor appeals in the context of reactance. Finally, I present the specific hypotheses and research questions associated with each of this study's four aims.

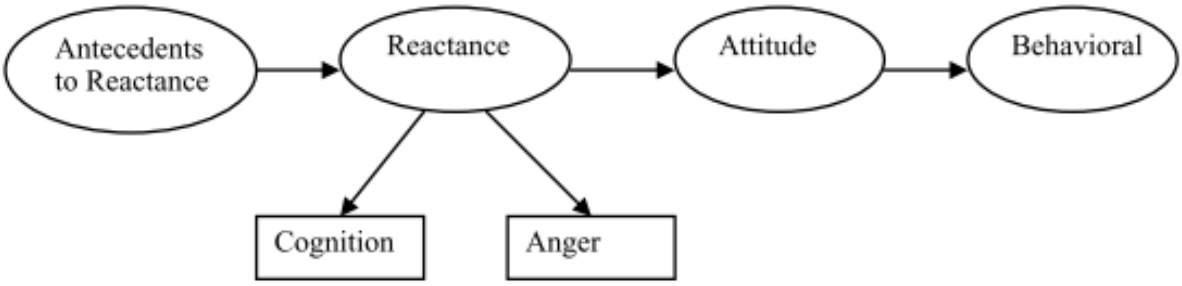
Psychological reactance theory and the intertwined model of reactance

Coined by social psychologist Jack Brehm in 1966, the term *psychological reactance* describes a motivational state that people experience when they feel compelled to assert their freedom to choose after perceiving some external force as a threat to their autonomy (Brehm, 1966). According to PRT, this state drives the so-called "boomerang effect" sometimes elicited by persuasive appeals, which causes audiences to change their behaviors in ways that are opposite message designers' intentions (Hovland et al., 1953). In the context of mass media campaigns, reactance occurs when people perceive messages as manipulative or otherwise threatening to their freedom to choose, then adopt or increase message-inconsistent behaviors as a way of restoring their sense of freedom and autonomy (Brehm, 1966; 1989). People may also

restore their freedom in less direct ways, such as by changing their attitudes toward the behavior in question, rejecting a message's information about threats or negative consequences, or derogating the message or its source (Smith, 1977; Wicklund, 1974). In all cases, people first believe they have *freedom* to make behavioral choices, then come to believe that persuasive appeals constitute a *freedom threat* and are motivated by *psychological reactance* to engage in some kind of *freedom restoration* behavior.

The actual state of psychological reactance was long thought to elude measurement, being described by its originator as a hypothetical variable (Brehm, 1989) that could only be inferred by the observation of its cause (freedom threat) and its antecedent (freedom restoration). This notion was challenged by Dillard and Shen (2005), who suggested the construct could be effectively measured using a composite index of anger and negative cognitions. Frustrated by the "ephemeral nature" (p. 145) of the reactance construct as initially proposed, these authors offered an empirically supported *intertwined model of psychological reactance* (illustrated in figure 3.1 below) that suggests an approach for measuring reactance and describes the psychological and affective processes underlying PRT in more detail. This model has since been supported in a variety of contexts (e.g. Quick & Bates, 2010; Rains & Turner, 2007; Shen, 2011) and has been identified by meta-analyses (Rains, 2013) and reviews (Reynolds-Tylus, 2019; Steindl et al., 2015) as the most successful and influential model of the theory. Thus, the understanding of reactance offered in this chapter will examine the model proposed by Dillard and Shen (2005) rather than of PRT's original authors (Brehm, 1966; 1989)

Figure 3.1: The Intertwined Model of Reactance



Note: *Figure from Dillard & Shen, 2005*

Perhaps the most notable contribution of Dillard and Shen's (2005) intertwined model of reactance is their assertion that this phenomenon is both cognitive and affective. Using a series of mediation models, the authors tested four different models of the reactance process: one only cognitive, one only affective, one in which cognitive and affective processes operated in parallel, and one in which cognitive and affective processes were *intertwined* and both indicated a single latent reactance factor. They found that the latter of these models was the best fit for their data, and this model has also proven superior in subsequent studies (Rains, 2013), though most evidence is consistent with both intertwined and dual-process models. In keeping with the understanding of reactance as an aversive motivational process (Brehm, 1989), they proposed that the valences of both cognitive and affective reactance processes are negative.

A meta-analysis of the reactance literature (Rains, 2013) confirmed that the intertwined model of reactance remains the best fit for data in studies of reactance, though the author of that analysis cautioned that the correlation between negative cognitions and anger was not as strong as one might expect. Still, the body of available evidence suggests that reactance can best be understood as either a combination of anger and negative cognitions or as some latent process that is otherwise associated with those two experiences. However, this evidence still relies on

introspection and providing a subjective interpretation of their experience. Thus, the current study aims to objectively examine the underlying neural processes during message exposure to determine whether there is support for the intertwined model of reactance.

Utilizing neural synchrony to identify message effects

Examining the brain through neuroimaging tools like functional magnetic resonance imaging (fMRI) allow researchers to observe processes like reactance in real time and without need for participant introspection, and this has been shown to extend our theoretical understanding of a number of message effects (Weber et al., 2018). When examining highly complex stimuli like videos, researchers often do not focus on average or mean level measures of activation alone but instead focus on measures of neural *synchrony* between audience members that account for dynamic nature of activation (i.e. patterns of increased and decreased brain activity that occur over the course of the stimuli presentation) in the brain.

When participants undergoing fMRI are exposed to the same dynamic stimuli, the patterns of activity within their brains will sometimes "sync up" to those stimuli in a way that is common across participants (e.g. Hasson et al., 2004; 2008). Even when participants experience the stimulus at independent times, the ebb and flow of blood throughout their brains tends to follow a similar rhythm during stimulus exposure. This phenomenon can be measured using intersubject correlation (ISC) analysis, which quantifies this similarity by comparing time series of neural activation across participants at the voxel level. This can be done by computing pairwise correlations for the time course of each voxel in template space or by correlating the voxelwise time course for each participant against an average time course for the full sample. The values for each pair or participant can then be incorporated into group-level analyses using a variety of methods (for a complete overview, see Natase et al., 2019). This enables researchers to

construct a number of useful measures, the simplest of which is a single coefficient for the group-level ISC for each voxel in the template brain. As with general linear model analyses like the one used in the preceding chapter, these values can also be averaged over regions or networks.

Importantly, ISC has been shown to extend to larger portions of the brain as a factor of audience engagement with the stimuli. For example, Hasson et al. (2008) found that ISC was mainly constrained to visual/auditory processing regions when participants watched raw footage of people milling about in a park but extended to regions associated with higher-order processing when those same participants watched a clip from Sergio Leone's *The Good, the Bad, and the Ugly*. When calculating the proportion of the cortex in which they observed ISC during exposure to the films, these authors found that this proportion varied as a factor of the films' editing and narrative structure. While Leone's classic elicited ISC in about 45% of the cortex, a film by master Alfred Hitchcock elicited ISC in more than 65% of the cortex. By contrast, a clip from the relatively unstructured show *Curb Your Enthusiasm* elicited ISC in less than 20% of the cortex and entirely unstructured footage of people milling about in a park less than 5%. These examples highlight how effective media can induce similar cognitive processes across audience members.

Research also suggests that the nature of processes "controlled" by stimuli can be inferred by the regions of the brain in which they are observed. For example, when Schmälzle and Grall (2020) showed participants a suspenseful film clip, they observed ISC in brain regions that are often associated with saliency and executive control. Further, they found that ISC signal in these regions was higher during portions of the film that an independent sample had indicated were more suspenseful. Similarly, when Imhof et al. (2020) presented participants with both

effective and anti-effective anti-alcohol PSAs, they found that effective ads elicited more ISC in the dorsomedial prefrontal cortex (dmPFC) and the precuneus, both areas associated with perceptions of personal relevance and involvement. Thus, the location of ISC signals during encoding of complex messages seems to indicate the cognitive and affective processes those messages elicit.

In sum, researchers use ISC to identify which regions of the brain are engaged by complex stimuli like PSAs. This measure indicates a common signal that is produced by the message and is shared across participants, screening out idiosyncratic responses that are unique to individuals. Further, past work has established that the processes indicated by this signal can sometimes also be detected using different methods in independent samples, providing an opportunity to triangulate the nature of the process itself. In the sections below, we identify three regions that have been identified with cognitive or affective processes that resemble three key components of the intertwined model of reactance: freedom threat perceptions, anger, and negative cognitions.

Freedom threat perceptions – Temporoparietal junction

Assuming that people already believe they possess a specific freedom, the process of message rejection via reactance begins with the perception that a message constitutes or contains a threat to that freedom. Freedom threats in the context of communication research are "most often attempts at social influence" (p. 2, Reynolds-Tylus, 2019), and experimenters often manipulate the strength of freedom threats by including "controlling" (p. 222, Miller et al., 2007) or "dogmatic" (p.450, Quick & Stephenson, 2008) language. For example, Worchel and Brehm (1970) created their "strong" freedom threat manipulation by adding phrases like "you cannot believe otherwise" and "you have no choice but to believe this" to a persuasive appeal. The

freedom threat perception itself is most often measured using Dillard and Shen's (2005) scale, which contains likert-style items such as "the message tried to pressure me", and "the message tried to manipulate me."

The wording of these measures suggests that freedom threat perceptions can be understood as perceptions of unwelcome attempts at social influence. Further, measures of that perception use statements about the messages' intentions, suggesting that the freedom threat perceptions thought to cause reactance are essentially inferences about the intentions that underlie these messages. Thus, if freedom threat perception involves making quick inferences about the intentions of another it would suggest that this process involves neural processes associated with mentalizing (Van Overwalle & Baetens, 2009) or considering the mental states of others.

The process of making sense of the contents of others' minds has been broadly associated with the mPFC, the temporoparietal junction (TPJ) and the precuneus (Van Overwalle & Baetens, 2009), as well as the posterior cingulate cortex (PCC; Saxe & Kanwisher, 2003; Decety & Lamm, 2007). In a meta-analysis of more than 200 mentalizing studies, Van Overwalle and Beatens (2009) found that the TPJ was the region most reliably associated with inferences about the intentions or goals of others. Comparing the TPJ to the mPFC and precuneus, these authors concluded that "the TPJ is the core area in goal inferences" (p. 578) and suggested that "early and automatic goal inferences always involve the TPJ." (p. 589). A later meta-analysis by Schurz et al. (2014) found that portions of the TPJ were activated in response to all theory of mind tasks. Thus, the TPJ is arguably the brain region most reliably associated with the process of mentalizing and it is associated with the specific aspect of mentalizing thought to be involved in freedom threat detection: inferences about the goals or intentions of another.

The term "temporoparietal junction" describes the region that connects the brain's temporal and parietal lobes on either side of the brain. Thus, the TPJ is actually two separate regions—one on the right side of the brain and one on the left. The meta-analyses above implicate the bilateral TPJ (i.e. both regions) in mentalizing, but there is also evidence implicating each individual side in the processing of harmful or deceitful intentions. Harada et al. (2009) observed bilateral TPJ activation when they asked participants to judge whether a fictional character was lying, and they found that the left TPJ was more sensitive to lies about antisocial behaviors than about prosocial behaviors. On the other side of the brain, when Young, Dodell-Feder, & Saxe (2010) asked participants to consider the moral implications of a violent act, they found that participants whose right TPJ had been disrupted with a magnetic field were less likely to consider the intentions of the violent actor in their judgment.

In sum, activation of the bilateral TPJ is reliably associated with inferences about the intentions and goals of others. Further, each TPJ region has been associated with judgments about others' intentions in the context of harm or deceit. These responses have been described as "early and automatic" (p. 589, Van Overwalle & Baetens, 2009), suggesting they are more associated with quick heuristic assessments than deep cognitions. These features align with descriptions of freedom threat perception, which is presented in the literature as a morally-valenced (negative) inference about the intentions of another that precedes an emotional response. Thus, we propose that freedom threat perceptions will be associated with activation of one or both TPJ regions.

Anger - Anterior Insula

Dillard and Shen operationalized the affective component of reactance as *anger*, reasoning that this "align[ed] well with Brehm's description of reactance as the experience of

hostile and aggressive feelings." (p. 147). Indeed, emotional descriptions of reactance often suggest the presence of anger. Brehm (1981) wrote that the experience of reactance "may be accompanied by feelings of hostility," (p.392), and Brehm has written that people experiencing reactance will "raise hell" and "will be throwing tantrums" in response to freedom threats (Brehm, 1989). Dillard and Shen measured anger by asking participants to self-report their experiences of related feelings (e.g. annoyance, irritation) using semantic differential scales ("[0 = none of this feeling] to [4 = a great deal of this feeling]"), and this procedure has been followed many times since (Rains, 2013; Reynolds-Tylus, 2019).

Neuroscientists seeking to observe anger processes in the brain often manipulate anger in one of two ways: either they show participants images of people demonstrating anger (e.g. angry faces), or they induce feelings of anger with experiences like unfair consequences in games (e.g. Feng, Luo, & Krueger, 2015). In a recent meta-analysis of these studies, Sorella et al. (2021) found that when researchers experimentally induced anger, they usually observed activity in the AI and the ventrolateral prefrontal cortex (vlPFC). Given past research examining the functions of these structures, these authors suggested that vlPFC in this context could be "involved in the evaluation of the emotion and in the modulation of the behavioral outcomes" (p. 9), while activity in the AI was likely related to interoception and the subjective experience of the emotion itself. This latter suggestion—that the subjective experience of anger is associated with activity in the AI—has been echoed in several other reviews and meta-analyses. A review by Alia-Klein et al. (2020) found that experimenter-induced anger was usually linked to activation of the AI, thalamus, and amygdala, and suggested that "these brain regions facilitate autonomic arousal, interoception and activation of the stress response" (p. 486). Similarly, Gilam and Hendler's (2017) review of human and rodent studies of anger noted that activation of the AI and dorsal

anterior cingulate (dACC) were recurrent observations in these studies. Again, these authors suggested that these regions were implicated in distinct processes, with the AI being associated with the emotional experience itself and the dACC facilitating behavioral control. Indeed, when Lindquist et al. (2012) concluded after a large meta-analysis that discrete emotions are not reliably associated with specific brain regions, they acknowledged the unusual finding that anger was uniquely associated with a pair of voxels within the AI. Thus, we argue that the most reliable neural index for the experience of anger is activity within the AI.

Further, many experiments have observed AI activation when participants are motivated by anger or indignation to perform a behavior that runs contrary to their interests. In ultimatum game paradigms, participants are asked to play a game in which a (usually artificial) co-player makes them offers to split sums of money. The participant is told that both players will receive the money offered if they accept but neither participant will receive any money if they reject. Thus, it is always in participants' best interest to accept offers—but participants often reject offers they believe to be unfair, incurring a personal cost. A meta-analysis of fMRI studies that used this paradigm (Feng, Luo, & Krueger, 2015) found that AI activation in response to unfair offers predicted the likelihood that these offers would be rejected and suggested that the AI might be part of a network associated with "reflexive and intuitive responses to norm violations, representing a motivation to punish the violators." (p. 599). Though PRT does not propose any motivation for punishment, this language does align with PRT's conceptualization of reactance as a motivational state induced by an unwelcome attempt at social influence and leading to a behavior that is opposed to the one advocated by that social influence. Thus, we suggest that the anger thought to indicate reactance will be associated with activity within the AI.

Negative cognitions – dorsolateral prefrontal cortex

Dillard and Shen (2005) measured the cognitive components of reactance by asking participants to list whatever thoughts they had about the message, then coding each thought as supportive, negative, or neutral. The number of negative-coded thoughts served as their measure of *negative cognitions*. In subsequent studies, researchers have had success with the less resource efficient procedure of asking participants to self-report the valence of their thoughts on semantic differential scales (e.g. "My thoughts about this PSA were … [1 = negative] to [7 = positive]" (Miller et al., 2007; Varava & Quick, 2015; Gardner & Leshner, 2016). In both cases, the cognitive component of reactance is understood as a negative valuation of the persuasive appeal.

Some recent research has examined the way message-induced negative cognitions manifest in the brain, but the patterns identified by that work so far vary depending on individual differences, message contexts, and other elements of each study. Weber et al. (2015) found that activity in parts the temporal lobe during counterarguing could predict acceptance of anti-drug message by high-risk individuals, while Liu et al. (2021) found that deliberative thinking and negative argumentation in response to messages were associated with different subregions of the dorsolateral prefrontal cortex (dlPFC), which had been identified as relevant to counterarguing in a localizer task (O'Donnell et al., 2018). Huskey et al. (2017) suggested that the networks involved in processing persuasive messages were moderated by interactions between message features and the audience's involvement with the message, partly because the certain interactions can drive some audience members to generate counterarguments. Thus, research in this area has not yet identified any neural correlates of message-induced negative cognitions that are thought to be generalizable across individuals and message contexts.

Though little research has tied the region directly to the processing of persuasive messages, there is other evidence to suggest the dlPFC might facilitate the cognitive component of reactance. Activity in the dlPFC has been associated with the processing of unfair offers in the ultimatum game (Sanfey et al., 2003; Speitel et al., 2019), but this region appears to play a different function than the AI. While AI activity scales with the degree of unfairness and is associated with rejection of unfair offers, the degree of dlPFC activity has been observed to not vary as a factor of unfairness (Sanfey et al., 2003). Players who have activity within this region disrupted using TMS are more likely to accept more unfair offers (Baumgartner et al., 2011; Knoch et al., 2006; 2008), and some have suggested this is because the dlPFC integrates with other frontal regions (i.e. vmPFC) to facilitate the decision to incur a personal cost in order to enforce norms of fairness (Baumgartner et al., 2011). Similarly, it has been hypothesized that the dlPFC exerts top-down control over the AI during the decision-making process (Sanfrey et al., 2003).

In sum, the bilateral dlPFC is among few brain regions that have been implicated in counterarguing to persuasive messages, and it is the only region to have been associated with counterarguing in a brain-as-predictor style study. The region has also been implicated in the cognitive component of a process that is similar to message rejection by reactance (rejection of unfair offers in the ultimatum game). Thus, we propose that the negative cognitions involved in reactance can be indexed by activation of the dlPFC.

Effects of fear and humor appeals on psychological reactance

One feature shared by many persuasive messages that is thought to elicit psychological reactance is the fear appeal (Shen, 2015). Fear appeal messages attempt to persuade audiences by arousing fear of negative outcomes that may befall them if they adopt, continue, or fail to adopt a

certain behavior. This approach is thought to be particularly prone to backfire (Finckenauer, 1982; Covello, Slovic, & Von Winterfeldt, 1986), partly because fear appeal messages are often seen by audiences as manipulative, creating freedom threat perceptions (Shen & Coles, 2015). Indeed, the extended parallel processing model (EPPM; Witte, 1992), which is among the most prominent theoretical frameworks for understanding fear appeals, specifically cites reactance as a cause of failure for fear-inducing messages. The tendency of fear appeals to induce reactance has also been observed empirically, with meta-analyses (Witte & Allen, 2000), experimental (Shen, 2011) and quasi-experimental (Shen & Coles, 2015) studies all showing that inducing fear in audiences can result in anger, negative cognitions, and freedom restoration behaviors (i.e. message derogation).

Humor appeals represent an emotional counterpoint to fear appeals—these messages are intended to induce positive affect in their audiences, producing positive attitudes toward the message and reducing negative cognitions and counterarguing (Makerjee & Dubé, 2012). This approach has become popular in the world of marketing and advertising (Eisend 2009; 2011), but little work has been done exploring the cognitive and affective mechanisms by which they are thought to cause persuasion (Nabi, 2015). Indeed, reviews and meta-analyses have observed no consistent persuasive effect of humor (Weinberger & Gulas, 1992), or have found that humor appeals are inferior to other emotional appeals (Lee & Cheng, 2010).

Overall, research suggests that humor appeal messages are less likely to induce reactance than fear appeal messages. Indeed, incorporating humor into fear appeal messages has been shown to mitigate the reactance caused by fear (Shen & Coles, 2014; Makerjee & Dubé, 2012). Positive emotions brought on by humor appeals have been associated with decreased reactance in a variety of contexts, including anti-alcohol messages (Sklaski et al., 2009), pro-vaccine

messages (Moyer-Gusé, Robinson, & Mcknight, 2019), and messages warning against purchasing unverified prescription medicines online (Alhabash et al., 2022). Thus, there is considerable evidence to suggest that fear/humor appeal is an important message feature with regards to reactance.

Objective, aims, and hypotheses

The objective of this study was to improve the theoretical understanding of both anti-DUIC messages and psychological reactance by connecting neural activation during the encoding of anti-DUIC PSAs with levels of self-reported reactance in an independent sample. Specifically, this study had four aims:

Aim 1: Establish that anti-DUIC PSAs elicit reactance and vary in reactance

Given that reactance has not yet been established as an important predictor of responses to anti-DUIC messages, we sought to establish that a prominent model of reactance (Dillard & Shen, 2005) would hold in this context. Further, establishing that our anti-DUIC stimuli could induce some degree of reactance was an important prerequisite for our attempts to identify neural correlates of reactance.

> *H1: The structure of survey data will align with predictions made by the intertwined model of reactance, such that reactance can be assessed as a higher-order latent variable indicated by two latent factors: anger and negative cognitions*
>
> OR
>
> *$H1_{alt}$: The structure of survey data will align with predictions made by the dual-process model of reactance, such that reactance describes parallel processes of anger and negative cognitions.*

Second, because past theory suggests that psychological reactance is produced by elements of persuasive messages themselves, we expect that the stimuli used in this study will elicit different levels of reactance.

H2: Values of the latent reactance or variable will vary across stimuli.

Aim 2: Determine whether anti-DUIC PSAs synchronize neural activity

The second aim of this study was to test whether neural responses to our stimuli could be used to predict self-reported reactance processes elicited by those same stimuli in an independent survey sample. Naturalistic stimuli like videos align activity across multiple participants' brains (Hasson, 2008), and more engaging stimuli align activity across larger portions of the brain.

H3: Professionally-produced anti-DUIC PSAs will align activity across a greater portion of the cortex than unstructured control stimuli.

Because past research suggests that effective PSAs elicit more ISC than ineffective PSAs (Imhof et al., 2020;), we also expect that PSAs rated as more effective by our survey sample will elicit more ISC than messages rated as less effective.

H4: PSAs that elicit more ISC across the cortex in the scanner sample will also elicit higher PME ratings from the survey sample.

Aim 3: Identify neural correlates of reactance processes

The third aim of this study was to test whether activity in certain brain regions during the encoding of anti-DUIC PSAs could be used to predict which messages induced stronger reactance processes than others in an independent sample. Past research has found that neural synchrony elicited by naturalistic stimuli is localized to brain regions involved in processing those stimuli, so we hypothesize that messages that elicit more self-reported reactance processes

from survey participants will also elicit more synchrony in brain regions likely to be implicated in those processes.

H5a: PSAs that elicit more ISC within the TPJ in the scanner sample will also elicit higher self-reported freedom threat perceptions from the survey sample.

H5b: PSAs that elicit more ISC within the AI in the scanner sample will also elicit higher self-reported anger from the survey sample.

H5c: PSAs that elicit more ISC within the dlPFC in the scanner sample will also elicit higher self-reported negative cognitions in the survey sample.

Aim 4: Test the effects of appeal type on reactance processes

The final aim of this study was to test whether the degree to which participants experienced reactance while viewing PSAs was affected by whether the messages were designed to use humor appeals or fear appeals. Past research suggests that messages using humor appeals elicit lower levels of reactance than messages that use fear appeals (Alhabash et al., 2022, Moyer-Gusé, Robinson, & Mcknight, 2019, Sklaski et al., 2009), so we hypothesize that both self-reported measures of reactance processes and their hypothesized neural correlates will be sensitive to the type of appeal used in each message.

H6a: PSAs that used humor appeals will elicit lower levels of self-reported anger in the survey sample than PSAs that used fear appeals.

H6b: PSAs that used humor appeals will elicit less ISC within the AI in the scanner sample than PSAs that used fear appeals.

H7a: PSAs that used humor appeals will elicit lower levels of self-reported negative cognitions in the survey sample than PSAs that used fear appeals

H7b: PSAs that used humor appeals will elicit less ISC within the dlPFC in the scanner sample than PSAs that used fear appeals.

Methods

This study was designed to improve understanding of the mechanisms by which messages elicit experiences of reactance in their audiences. To ensure that the effects we observed could be attributed to the messages themselves, we presented the same set of messages to two independent samples using two different data collection protocols: an online survey and an fMRI procedure.

Participants

Online survey

Participants for the online survey portion of this study (N = 236) were recruited using the online recruitment service Qualtrics Panels. All participants were between 18 and 24 years of age (M = 21.77, SD = 1.75) and had scored at least a 2 on the Cannabis Use Disorder Identification Test (CUDIT). Participants scored an average of 12.46 on this measure (SD = 6.82), and 74.15% of participants reported a score that exceeded the threshold for CUD (a score of 8 or more). The sample was mostly predominantly white (61.44%) and female (52.97%).

fMRI protocol

Participants for the neural portion of this study were recruited using an online job board hosted by a large midwestern research university. A total of 40 participants underwent the scanning procedure, but due to software and mechanical errors only 29 participants viewed the full set of PSA stimuli. Thus the sample retained for analysis was N = 29. All participants were between 18 and 24 years of age (M = 19.5, SD = 1.45) and scored at least a 2 on the Cannabis Use Disorder Identification Test (CUDIT). Participants scored an average of 16.24 on this measure (SD = 45.28), and 100% exceeded the threshold for CUD. The sample was predominantly white (72.41%) and female (55.17%). All participants were right-handed, had

normal (or corrected to normal) vision, were not taking any psychoactive medications, did not suffer from claustrophobia, and did not have metal in their bodies that was contraindicated for MRI.

Procedures

Online survey

After being screened, participants viewed a series of eight 30-second anti-DUIC PSAs, presented in a random order. All PSAs were professionally produced and distributed in major North American markets (seven in the United States and one in Canada), and all explicitly advocated against cannabis-impaired driving. Four of the messages used fear appeals and four used humor appeals. After viewing each message, participants were presented a series of self-report measures of PRT constructs. To mimic the order of events described by the intertwined and dual-process models of reactance (Dillard & Shen, 2005), items were presented in the order below. Because both models describe anger and negative cognitions as being experienced simultaneously, these two measures were presented in a counterbalanced random order.

Freedom threat perceptions

In the online survey, freedom threat perceptions were measured using a slight modification of the scale of four Likert-style items proposed by Dillard and Shen (2005). After feedback from test participants, wording of these items was adjusted slightly so that the agent of the freedom threat became the *source* of the message rather than the message itself. After exposure to each PSA, participants were asked to rate the degree to which they agreed with the following statements on a scale of 1 (strongly disagree) to 5 (strongly agree). Items were presented in a random order.

1. The makers of this PSA were trying to make my choices for me.
2. The makers of this PSA don't respect my freedom to choose.
3. The makers of this PSA were trying to manipulate me.
4. The makers of this PSA were trying to pressure me.

Anger

In the online survey, message-elicited anger was measured by averaging responses to self-reports of two emotions: anger and annoyance. After exposure to each PSA, participants were asked to indicate the degree to which they agreed or disagreed with the following statements on a scale of 1(strongly disagree) to 7 (strongly agree). Statements were presented in a random order.

The PSA made me feel…

1. Angry
2. Annoyed
3. Bored
4. Amused
5. Interested
6. Anxious
7. Afraid

Values from the first two items were averaged to create an index of anger.

Negative cognitions

Negative cognitions were assessed by averaging responses to three semantic differential items participants were asked to respond to the following statements on a scale of 1(left option) to 7 (right option). Statements were presented in a random order.

My thoughts toward this PSA are mostly…

1. Bad/Good
2. Unfavorable/Favorable
3. Negative/Positive

Freedom restoration

Because this procedure did not allow us to observe actual behaviors, freedom restoration was measured using an accepted attitudinal proxy: message derogation. Message derogation was measured using four items that comprise a "derogation" factor in the larger Reactions to Health Warnings Scale (Hall et al., 2017), which is designed to capture reactance and other negative responses to messages about health risks. Participants were asked to indicate the degree to which they agreed with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree). Statements were presented in a random order.

1. This PSA's message was exaggerated
2. This PSA's message was distorted
3. This PSA's message was misleading
4. This PSA's message was overblown

fMRI protocol

In the neural sample, participants viewed 22 randomly ordered 30-second videos while undergoing fMRI across two consecutive eight-minute runs. After viewing each video, participants responded to the prompt "How effective was this PSA?" to rate their perceptions of the message's effectiveness (PME) on a scale of one to four using a four-button Current Designs response pad held in the participant's right hand. Participants were given three seconds to enter a rating, then viewed a three-second countdown before presentation of the next PSA. To ensure that only steady-state images were used in analysis, presentation of videos did not begin until after eight seconds of scan time.

Participants viewed three different types of clips during the fMRI procedure: five clips contained unstructured control footage (pedestrians on city streets), seven advocated against texting and driving, and 12 advocated against DUIC. Of the 12 anti-DUIC messages, six used fear or demonstrated negative consequences and five were classified as "humor" appeals because

they used a lighthearted, humorous tone. One did not fit into either category and was instead classified as a "moral appeal" message. This sample included all PSAs included in the online survey portion of this study.

Due to problems during data collection, not all participants in the study viewed all messages in the sample. Only messages that were shown to our complete subsample of participants were selected for analysis, so our final set of stimuli consisted of the eight messages included in the online survey (four fear, four humor) and three control clips of unstructured footage.

fMRI data acquisition

Structural and functional brain imaging was conducted using a 3 Tesla GE Discovery MR750 scanner. Head motion was minimized with foam padding on the head coil. The structural imaging procedure was changed mid-study to reduce scan time. Anatomical images were obtained using a T1-weighted FSPGR BRAVO sequence. Two functional runs were recorded (TR = 800ms, TE = 20ms, flip angle = 60°, matrix size = 96x96, 54 axial slices, 3mm thick; voxel size = 3.0x.2.5x2.5). PSAs were projected onto a screen on the back of the scanner bore, which participants viewed through a mirror mounted directly above the head coil. Participants heard sound through MR-compatible headphones designed to operate within the bore.

Preprocessing was performed using the afni_proc.py program within the Analysis for Functional Neuroimaging (AFNI) software package (Cox, 1996). In this process, slice timing correction was performed using the first slice as reference, and images were corrected for minor motion by spatially realigning all images to the first acquired volume. Structural images were registered to the main functional image and skull-stripped. Functional and structural. Functional

and structural runs were warped to align with the Montreal Neurological Institute MNI151 template brain and smoothed with a 4mm Gaussian kernel.

A Priori Regions of Interest

Using past research as a guide, we constructed neural masks for each of our *a priori* regions of interest. We also created masks for three regions for use as controls: two regions that have been associated with positive valuation and reward processes that have the opposite affective valence as reactance (Ventral Striatum and vmPFC), and a portion of the Visual Cortex that is not thought to be associated with higher-order processing. Coordinates of all ROIs are listed in table 3.2 below, and ROIs associated with hypotheses are illustrated in figure 3.2[1].

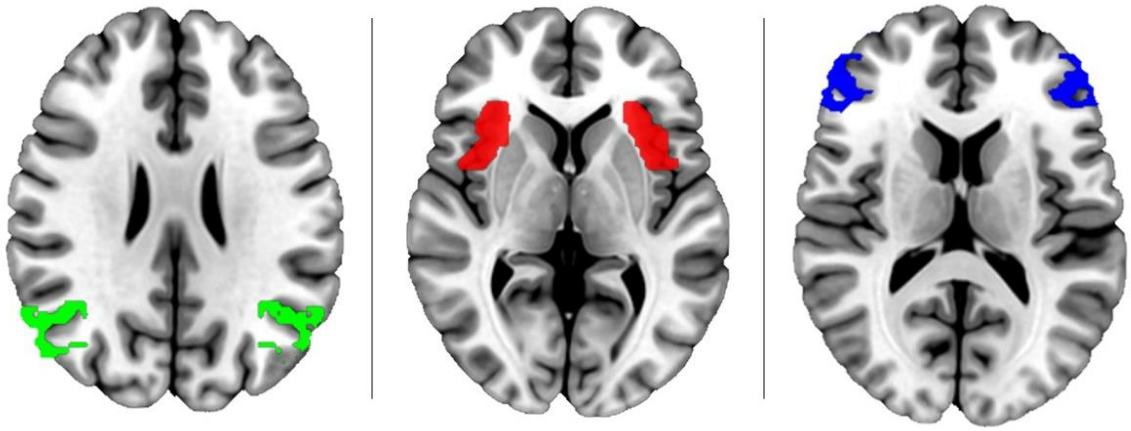
Table 3.2: A Priori Regions of Interest

	X	*Y*	*Z*	*Size (voxels)*
TPJ	55	69.2	18	281
	-52.5	69.2	18	281
AI	-32.5	-10.8	-24.5	925
	35.0	-13.2	-24.5	771
dlPFC	-50.0	-38.2	0.5	275
	50.0	-38.2	0.5	272
VS	-10.0	-8.2	-12.0	235
vmPFC	2.5	-35.8	-17	233
Visual Cortex	20.0	96.0	-19.5	1,734

Note: *X, Y, and Z coordinates are presented in the MNI template brain*

[1] Axial, sagittal, and coronal slices illustrating all six ROIs can be found in the supplementary materials for Chapter 3

Figure 3.2: A Priori Regions of Interest



Note: *Axial slices illustrating all three a priori ROIs: TPJ (left), AI (middle), and dlPFC (right*

Analysis

Aim 1: Establish that anti-DUIC PSAs elicit reactance and vary in reactance

Before comparing our two datasets, we sought to confirm that the stimuli used in this study induced varying degrees of reactance. We tested whether the structure of our data aligned with established theory by comparing two models of reactance proposed by Dillard and Shen (2005): a dual process model in which the effects of freedom threat on attitudes are mediated by both anger and negative cognitions independently and an intertwined model in which this effect is mediated by a higher-order latent reactance factor, which is indicated by anger and negative cognitions.

We used lavaan (Rosseel, 2012) to test the fit of both models, modeling all variables as latent factors indicated by their respective self-report items. To account for the clustered nature of the data, we also modeled by-participant random intercepts. Fit indices were estimated using a robust maximum likelihood procedure and models were compared using the Bayesian Information Criterion.

The final step in our analysis depended on the results of our model comparison. If our comparison found that the intertwined model of reactance fit our data better, we would calculate an aggregate reactance score from our measures of anger and negative cognitions, then assess whether reactance scores differed as a factor of stimuli by testing the fit of a number of linear mixed effects models with the lme4 package in R. Each model would contain dummy-coded regressors for each stimulus as fixed effects and a by-participant random intercept and a by-participant random slope for each fixed effect. We would the fit of seven such models, allowing for pairwise comparisons of the reactance elicited by each set of stimuli.

If our dual-process model showed better model fit, we would conduct the same procedure on anger and negative cognition values independently.

Aim 2: Determine whether anti-DUIC PSAs synchronize neural activity

To test whether our PSA stimuli synchronized participant's brain activity more than unstructured control messages, we used an inter-subject correlation (ISC) analysis. First we extracted the images obtained during the encoding of each individual 30-second message, producing a total of 11 images for each participant (one for each of the eight PSAs and the three control messages). For each message, we then computed voxelwise Pearson correlation coefficients for all possible pairings of participants using the 3dTcorrelate command in AFNI. Correlation values were then subjected to a Fisher-Z transformation. Group-level voxelwise correlation values were estimated using linear mixed effects modeling with a cross random effects structure to account for the nested nature of the data (within individual participants) using the 3dISC program in AFNI (Chen et al., 2017). This process produced a single group-level image for each PSA with a value for each voxel indicating the degree to which variation in the activation of that voxel over the time series was shared among all participants.

To confirm that our anti-DUIC PSA stimuli aligned neural activity across a greater portion of participants' brains than our unstructured control videos, we corrected each sample-wide ISC map for multiple comparisons using a voxelwise false discovery rate threshold of $q < 0.05$ (Benjamini & Hochberg, 1995). We then divided the number of voxels in which the signal surpassed this threshold from the number of voxels in a mask of the entire cortex, producing a value of the percentage of the cortex that showed ISC in response to each message. We also visually inspected each ISC map to identify which brain regions were "controlled" by the corresponding message.

To test our hypotheses that PSAs that elicited more ISC across the cortex were also rated as more effective in our survey sample, we used a procedure that compared rank-orderings of PSAs by these two outcomes. This approach is typical of brain-as-predictor studies (e.g. Falk, Berkman, & Lieberman, 2012; Weber et al., 2015; Scholz et al., 2017) because the use of rank-orders allows for straightforward and valid comparisons across independent datasets using different styles of measurement.

This procedure consisted of the following steps. First, we produced a single vector representing the ranks of each of our eight PSAs by the percentage of the cortex that showed significant levels of ISC. Second, we estimated a vector for each survey participant representing the ranks of each PSA by the PME score the participant assigned to that PSA. We then computed Kendall rank correlation coefficients between each participant's rank vector and the ISC rank vector. This provided us with a value for each participant that quantified the degree to which their rank-ordering of PSAs by self-reported PME aligned with the rank-ordering of PSAs by shared neural signal in the cortexes of our scanner sample. Because this test recognizes comparisons between rank-orderings that only contain a single ranking (i.e. instances when

participant's issued the same self-report values in response to all PSAs) as perfect monotonic relationships, participant rank-orders that ranked all PSAs equally were excluded from the analysis. Finally, we averaged all coefficients to produce a single value for each sample-wide comparison.

To determine whether these values were significantly different from values that would be expected from chance, we conducted a Monte Carlo simulation resembling the procedure above. This procedure produced datasets with pseudo-random values for each participant's rating of each PSA that aligned with the range of values possible in the dataset (1-4), then produced a rank-ordering of these values for each participant. Again, a Kendall coefficient was computed comparing each rank-ordering to a fixed rank-ordering, and these coefficients were averaged for each dataset. We conducted 10,000 iterations of this procedure, providing us with distributions of 10,000 mean coefficients that would be expected by random ratings of the PSAs. We then estimated *p*-values for our Kendall correlation coefficient by comparing this statistic against the simulated distribution.

Aim 3: To identify neural correlates of reactance processes

To identify whether rank-orderings of PSAs by self-reported reactance experiences aligned with rank-orderings of PSAs by alignment in specific brain regions, we conducted a modified version of the procedure described for PME above for each combination of self-report variable and ROI (freedom threat and TPJ, anger and AI, negative cognitions and dlPFC). We also tested the rank-orderings by each self-report variable against rankings by ISC in three control ROIS: the ventral striatum (VS), the ventromedial prefrontal cortex (vmPFC), and the visual cortex.

Aim 4: Test the effects of appeal type on reactance processes

To test whether fear-themed PSAs elicited more self-reported anger and negative cognitions from our survey sample than humor-themed PSAs, we tested linear mixed-effects models in which anger/negative cognitions served as the outcome variable and appeal condition (fear/humor) serves as the predictor variable. To account for the nested nature of our data, we also modeled a by-participant random slope and intercept and a by-PSA random intercept.

To test whether fear-themed PSAs elicited more ISC in the AI and dlPFC, we compared the means of the ROI-wide ISC values across humor messages and fear messages using a Wilcoxon rank-sum test.

Results

Survey data align with a dual processing model of reactance

Indices suggested good model fit for both the intertwined $\chi^2(61) = 100.273$, $p = 0.001$, RMSEA = 0.019 90%CI[0.014, 0.025], $SRMR_{within} = 0.027$, $SRMR_{between} = 0.005$, BIC = 75,255.067 and dual process $\chi^2(61) = 124.433$, $p < 0.001$, RMSEA = 0.025 90%CI[0.020, 0.030], $SRMR_{within} = 0.034$, $SRMR_{between} = 0.007$, BIC = 74,863.968 models of reactance. All factor loadings and regression coefficients were significant in both models, but a comparison of the Bayesian Information Criterion suggested that the dual process model was a better fit for the data. Given this result, we tested two series of linear mixed-effects models to determine how levels of anger and negative cognitions differed across each possible pair of stimuli. As shown in Table 3.2, we observed significant differences in levels of negative cognitions across 18 of 28 possible pairings. Sixteen of these differences remained significant after correcting for multiple comparisons using the Holm-Bonferroni method. We observed significant differences in anger across 20 pairings, and all of these survived our multiple comparison correction (see Table 3.3).

Table 3.3: Pairwise Comparisons of Negative Cognitions

	vs. Rollover	Vs. Basketball	Vs. Party kids	Vs. Pineapple	Vs. Facemask	Vs. Canada	Vs. TV Mount
Basketball	β = -0.203 SE = 0.109 F(1,360.03) = 3.460 *p* =0.064 p_{adjust} = 0.540						
Party kids	β = 0.012 SE = 0.114 F(1,364.11) = 0.011 *p* = 0.918 p_{adjust} = 1.000	β = 0.214 SE = 0.112 F(1,357.37) = 3.687 *p* = 0.056 p_{adjust} = 0.540	--				
Pineapple	β = -0.741 SE = 0.108 F(1, 358.00) = 46.953 *p* < 0.001*** p_{adjust} < 0.001***	β = -0.538 SE = 0.102 F(1,348.89) = 27.385 *p* < 0.001*** p_{adjust} < 0.001***	β = -0.752 SE = 0.108 F(1,357.18) = 48.424 *p* < 0.001*** p_{adjust} < 0.001***	--			
Facemask	β = -0.491 SE = 0.117 F(1,366.46) = 17.626 *p* < 0.001*** p_{adjust} = < 0.001***	β = -0.288 SE = 0.114 F(1,358.65) = 6.412 *p* = 0.012** p_{adjust} = 0.144	β = -0.502 SE = 0.117 F(1,366.05) = 18.452 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.250 SE = 0.111 F(1, 348.31) = 5.058 *p* = 0.025* p_{adjust} = 0.275	--		
Canada	β = 0.136 SE = 0.108 F(1,358.05) = 1.605 *p* = 0.206 p_{adjust} = 1.000	β = 0.339 SE = 0.105 F(1,353.28) = 10.401 *p* = 0.001** p_{adjust} = 0.014*	β = 0.125 SE = 0.108 F(1,357.26) = 1.346 *p* = 0.247 p_{adjust} = 1.000	β = 0.877 SE = 0.102 F(1, 345.45) = 73,330 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.627 SE = 0.109 F(1,360.44) = 32.914 *p* < 0.001*** p_{adjust} < 0.001***	--	
Tv Mount	β = -0.412 SE = 0.112 F(1,363.79) = 13.550 *p* < 0.001*** p_{adjust} = 0.004**	β = -0.210 SE = 0.109 F(1,356.80) = 3.732 *p* = 0.054 p_{adjust} = 0.540	β = -0.424 SE = 0.112 F(1,363.13) = 14.336 *p* < 0.001*** p_{adjust} = 0.003**	β = 0.329 SE = 0.106 F(1, 347.76) = 9.646 *p* = 0.002** p_{adjust} = 0.026*	β = 0.079 SE = 0.112 F(1,364.05) = 0.496 *p* = 0.482 p_{adjust} < 1.000	β = -0.549 SE = 0.106 F(1,465.87) = 24.383 *p* < 0.001*** p_{adjust} < 0.001***	--
Eyeball	β = 0.163 SE = 0.111 F(1,362.25) = 2.179 *p* = 0.141 p_{adjust} = 0.987	β = 0.366 SE = 0.109 F(1,357.05) = 11.290 *p* < 0.001*** p_{adjust} = 0.013*	β = 0.152 SE = 0.111 F(1,362.14) = 1.867 *p* = 0.173 p_{adjust} = 1.000	β = 0.904 SE = 0.106 F(1, 347.96) = 72.326 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.654 SE = 0.112 F(1,364.48) = 33.932 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.027 SE = 0.106 F(1,355.54) = 0.063 *p* = 0.801 p_{adjust} = 1.000	β = 0.575 SE = 0.111 F(1,360.88) = 27.060 *p* < 0.001*** p_{adjust} < 0.001***

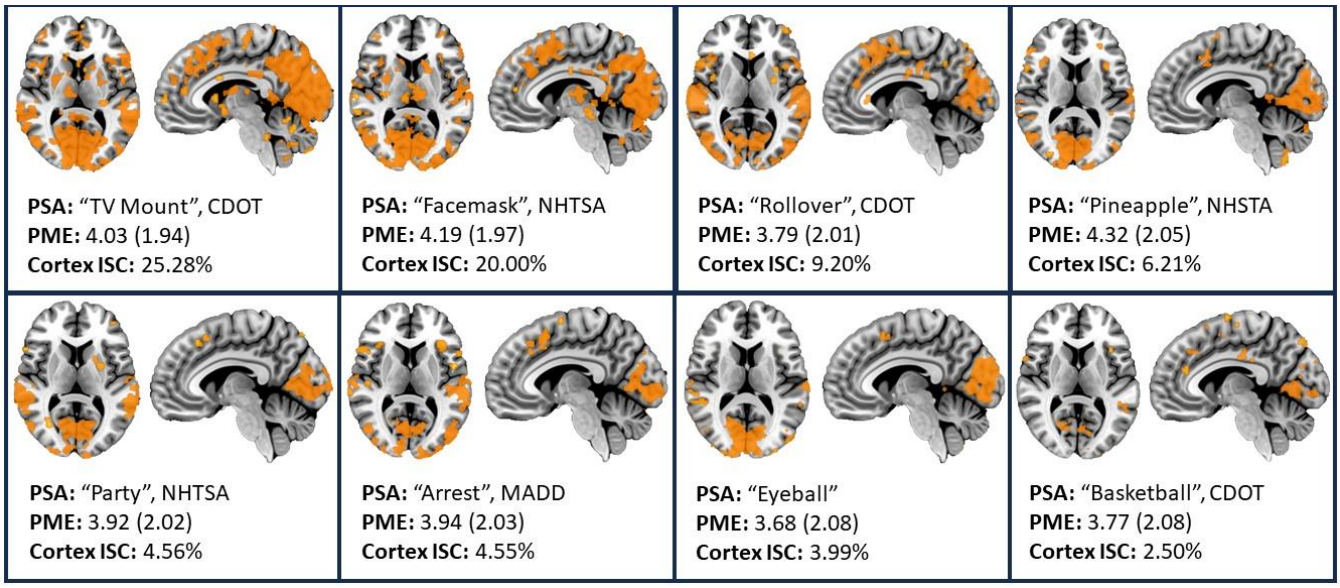
Table 3.4: Pairwise Comparisons of Anger

	vs. Rollover	Vs. Basketball	Vs. Party kids	Vs. Pineapple	Vs. Facemask	Vs. Canada	Vs. TV Mount
Basketball	β = -0.446 SE = 0.269 F(1,323.45) = 2.752 *p* = 0.098 p_{adjust} = 0.715						
Party kids	β = -0.006 SE = 0.252 F(1,323.59) = 0.001 *p* = 0.981 p_{adjust} = 1.000	β = 0.442 SE = 0.258 F(1,326.09) = 2.677 *p* = 0.103 p_{adjust} = 0.715	--				
Pineapple	β = -1.291 SE = 0.236 F(1,320.04) = 29.754 *p* < 0.001*** p_{adjust} < 0.001***	β = -0.874 SE = 0.246 F(1,323.36) = 12.572 *p* < 0.001*** p_{adjust} = 0.006**	β = -1.296 SE = 0.237 F(1,320.80) = 29.932 *p* < 0.001*** p_{adjust} < 0.001***	--			
Facemask	β = -1.215 SE = 0.240 F(1,319.53) = 25.615 *p* < 0.001*** p_{adjust} = < 0.001**	β = -0.796 SE = 0.250 F(1,322.99) = 10.156 *p* = 0.002** p_{adjust} = 0.017*	β = -1.218 SE = 0.237 F(1,321.68) = 25.459 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.081 SE = 0.229 F(1,313.67) = 0.125 *p* = 0.724 p_{adjust} = 1.000	--		
Canada	β = 0.809 SE = 0.261 F(1,327.23) = 9.563 *p* = 0.002** p_{adjust} = 0.022*	β = 1.233 SE = 0.266 F(1,329.41) = 21.407 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.808 SE = 0.261 F(1,327.40) = 9.550 *p* = 0.002** p_{adjust} = 0.022*	β = 2.104 SE = 0.251 F(1,316.17) = 70.220 *p* < 0.001*** p_{adjust} < 0.001***	β = 2.030 SE = 0.255 F(1,319.81) = 63.483 *p* < 0.001*** p_{adjust} < 0.001***	--	
Tv Mount	β = 0.983 SE = 0.252 F(1,323.83) = 15.226 *p* < 0.001*** p_{adjust} = 0.002**	β = 1.404 SE = 0.266 F(1,324.59) = 29.607 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.983 SE = 0.250 F(1,321.53) = 15.491 *p* < 0.001*** p_{adjust} = 0.002**	β = 2.278 SE = 0.235 F(1,314.03) = 93.777 *p* < 0.001*** p_{adjust} < 0.001***	β = 2.203 SE = 0.241 F(1,317.21) = 83.339 *p* < 0.001*** p_{adjust} < 0.001***	β = 0.172 SE = 0.258 F(1,419.93) = 0.443 *p* = 0.506 p_{adjust} = 1.000	--
Eyeball	β = 0.012 SE = 0.256 F(1,321.61) = 0.002 *p* = 0.962 p_{adjust} = 1.000	β = 0.444 SE = 0.260 F(1,324.16) = 2.902 *p* = 0.089 p_{adjust} = 0.715	β = 0.012 SE = 0.254 F(1, 320.96) = 0.002 *p* = 0.963 p_{adjust} = 1.000	β = 1.313 SE = 0.248 F(1,312.21) = 28.001 *p* < 0.001*** p_{adjust} < 0.001***	β = 1.228 SE = 0.253 F(1,317.76) = 23.583 *p* < 0.001*** p_{adjust} < 0.001***	β = -0.831 SE = 0.250 F(1,320.26) = 11.080 *p* = 0.001** p_{adjust} = 0.012*	β = -0.988 SE = 0.256 F(1,314.16) = 14.915 *p* < 0.001*** p_{adjust} = 0.002**

Anti-DUIC PSAs elicited similar patterns of ISC

Examination of the ISC maps for each of the eight anti-DUIC PSAs and three unstructured control messages confirmed that the structured PSAs aligned activity across greater portions of the brain. We found that the percentage of cortex in which we observed ISC varied widely across our PSA stimuli—the Colorado Department of Transportation's "Basketball" ad elicited ISC across just 2.50% of the cortex, while the same agency's "TV Mount" PSA elicited ISC across 25.28% of the cortex. By contrast, the best-performing control message elicited ISC across just 0.39% of the cortex, and the other two control messages showed no ISC signal after our correction for multiple comparisons. Percentages for each message are displayed in Figure 3.3 below.

Figure 3.3 – PSA ISC maps



Note: *Inter-subject correlation maps produced for each anti-DUIC PSA. Highlighting indicates voxels with correlation coefficients that exceeded FDR > .05 threshold. Images are ordered left to right in two rows by highest volume of ISC to lowest. PME indicates mean and standard deviation PME ratings for each PSA in survey sample.*

Visual inspection of the ISC maps produced for the eight PSAs suggested that all messages elicited ISC in certain brain regions. All maps showed large clusters of ISC in the occipital cortex as well as bilateral clusters in the temporal lobe, usually near the temporo-parietal junction. All maps also included a cluster of ISC in a caudal portion of the superior frontal gyrus. Though maps for messages that elicited large amounts of ISC also showed heterogenous distributions of clusters across the brain, the bulk of ISC in these maps seemed to radiate from the occipital and temporal-parietal clusters. Notably, three of the higher-ISC maps also showed clusters near the posterior cingulate and precuneus.

As hypothesized, we also found that participants' rank-orderings of PSAs by perceptions of message effectiveness were positively associated with a rank-ordering of PSAs by the extent of ISC within the cortex. In other words, the PSAs that elicited ISC across a larger portion of our scanner sample's brains also tended to be rated as more effective by our survey sample M_τ = 0.050 , p = 0.013.

Freedom threat was not associated with TPJ or control ROIs

Contrary to our hypothesis, we found that participants' rank-orderings of PSAs by perceptions of freedom threat were not associated with a rank-ordering of PSAs by the ISC they elicited in the temporal-parietal junction. In other words, the mean Kendall's Rank Coefficient for these by-participant comparisons was no different than would be expected under conditions of chance M_τ = -0.010 , p = 0.622. Rank-orderings by freedom threat perceptions were also not associated with rank-orderings by alignment in positive valuation regions (VS: M_τ = -0.013, p = 0.493; vmPFC: M_τ = -0.006, p = 0.772) or the visual cortex: M_τ = 0.026 , p = 0.192). Twenty-one survey participants' results were removed from this analysis because they offered identical Freedom Threat values for all eight PSAs.

PSAs that elicited more anger elicited more alignment in the AI, less in VS

Participants' rank-orderings of PSAs were positively associated with rank-orderings by ISC in the anterior insulae $M_\tau = 0.071$, $p < 0.001$ and negatively associated with activity in the ventral striatum $M_\tau = -0.059$, $p = 0.003$. We also observed a positive association with rank-orderings by ISC in the visual cortex $M_\tau = 0.091$, $p < 0.001$, but no association with rank-orderings by alignment in the vmPFC $M_\tau = -0.016$, $p = 0.091$. Two survey participants were excluded from this analysis because they offered identical Anger scores for all eight PSAs.

PSAs that elicited less negative cognitions elicited more alignment in the dlPFC

We did not observe a significant relationship between rank-orderings by negative cognitions and by ISC in the dorsolateral prefrontal cortex $M_\tau = -0.015$, $p = 0.442$, but we did observe a positive association with alignment in both value regions (VS: $M_\tau = 0.054$, $p = 0.006$; vmPFC: $M_\tau = 0.064$, $p = 0.002$). We did not observe a significant correlation with our rank-ordering by visual cortex activity $M_\tau = -0.026$, $p = 0.181$. Eighteen survey participants were excluded from this analysis because they offered identical Negative Cognition scores for all eight PSAs.

Fear appeal PSAs elicited more negative cognitions than humor appeal PSAs

Tests of our linear mixed-effects models found that levels of negative cognitions were affected by our fear/humor appeal condition, such that participants reported more negative cognitions while viewing fear appeal PSAs $\boldsymbol{\beta} = 0.540$, $F(1,9.115) = 16.672$, $p = 0.003$. The variance of our by-participant random intercept was 1.572 (SD = 1.254) and our by-participant random slope for appeal condition was 0.789 (SD = 0.888). The variance for this model's by-message random intercept was 0.023 (SD = 0.151).

Our message appeal condition did not have a significant effect on levels of anger **β** = 0.719, $F(1,6.125) = 1.563$, $p = 0.257$. The variance for our by-participant random intercept in this model was 7.048 (SD = 2.655), and for our by-participant random slope for appeal type was 0.797 (SD = 0.893). The variance for this model's by-message random intercept was 0.621 (SD = 0.788).

ROI activity was not affected by fear/humor appeals

Results of our Wilcoxon rank-sum tests indicated that ROI-wide ISC levels did not differ across fear/humor appeal PSAs in the AI $W = 14$, $p = 0.114$ or the dlPFC $W = 10$, $p = 0.686$).

Discussion

This study aimed to improve the understanding of message-induced psychological reactance by testing hypothesized neural correlates of important reactance constructs using an ISC approach. Specifically, we sought to triangulate a neural signal of reactance by testing whether anti-cannabis PSAs that aligned neural activity in certain regions of the brains of a small fMRI sample also elicited stronger self-reports of freedom threat perceptions, anger, and negative cognitions in an independent sample. Our results suggest that messages that align activity in certain brain regions may also induce higher levels of certain reactance experiences, but this interpretation is complicated by some unexpected findings.

Our results contribute to the understanding of both reactance and anti-DUIC messaging in three ways. First, they demonstrate that prominent models of reactance can be observed in the context of anti-DUIC messaging. Second, they show that professionally-produced anti-DUIC messages have the power to align neural activity in certain brain regions. Third, they suggest that

message-induced reactance may involve brain regions associated with social cognition and valuation.

Self-reported reactance to DUIC-messages aligned with theory-driven expectations

When participants in our survey sample were questioned about their experiences watching anti-DUIC PSAs, they reported experiences that aligned with both the intertwined and dual-process models of reactance. That is, participants tended to report higher levels of anger and more negative cognitions when viewing PSAs they felt posed a threat to their freedom to choose, and these experiences in turn predicted the generation of negative attitudes about the messages. The structure of this self-report data was consistent with the intertwined model of reactance, but a comparison of our structural equation models found it was better explained by the similar but distinct dual-process model. This differs slightly from past findings—Dillard and Shen (2005) first proposed the intertwined model after finding it outperformed a dual-process model, and Rains (2013) also found that the intertwined model was a better fit for data extracted from 20 past studies of reactance processes. In each of these cases, path coefficients were found to be significant for both models.

We tested the fit of these models for two reasons: to establish that the reactance processes we were interested in could be observed in response to our stimuli and to guide subsequent steps in our analysis. This study was not designed to compare the validity or utility of these models, and our findings suggest that both models fit our data to an acceptable degree. Thus, we do not assert that these findings suggest any weakness of the intertwined model of reactance or challenge to the understanding of reactance as a latent psychological construct. Instead, we view this as a likely idiosyncrasy within our dataset and see it most importantly as evidence that the relationships thought to define the reactance process were present in responses to these anti-

DUIC messages. Of particular interest to designers of anti-DUIC messages, this confirmed that audiences' negative attitudes toward these messages were directly associated with the degree to which they caused experiences of anger and negative cognitions. Specifically, negative cognitions had a small to moderate effect on negative attitudes ($\boldsymbol{\beta} = 0.244$), whereas anger had a moderate to large effect ($\boldsymbol{\beta} = 0.597$). Because our models included by-participant random effects, these effects cannot be attributed solely to differences in participants' baseline levels of anger or negative cognitions.

We also observed that some anti-DUIC PSAs reliably elicited more anger and negative cognitions than others, as evidenced by the fact that most of our pairwise comparisons of PSAs revealed significant differences in the mean values of these two variables. This finding supported our practice of rank-ordering PSAs by these values for comparison, and it demonstrated that anger and negative cognitions were caused at least partly by elements of the messages' design. This seemingly obvious finding is important because reactance experiences are thought to be partly responsible for the failure of past anti-cannabis campaigns (Hornik et al., 2008), and reactance to anti-cannabis messages has been shown to have "boomerang effects" in the form of increased cravings for and intention to use cannabis (Slavin & Earleywine, 2019). Given that established best practices for traffic safety campaigns hold that message content should be informed by theory (Wundersitz et al., 2010), these findings suggest that PRT may provide a useful theoretical framework for the design of anti-DUIC messages.

Anti-DUIC PSAs synchronize activity in specific brain regions

Just as the results from our survey showed that the content of anti-DUIC messages affected participants' self-reported levels of anger and negative cognitions, analysis of our fMRI messages revealed that these messages exerted control over the activity in certain regions of

participant's brains. The extent of this control varied widely across our eight PSAs, but we found that all PSAs showed increased ISC in the brain than our unstructured control footage. This is consistent with past research (Hasson et al., 2008) finding that structured, professionally-produced videos tend to show increased ISC in the cortex compared to control messages.

We found that all eight PSAs elicited relatively high levels of ISC in the occipital lobe and some degree of ISC in lateral portions of the temporal or parietal lobes near the temporal-parietal junction. These observations align with those of past studies using audiovisual stimuli (e.g. Hasson et al., 2004, 2008; Schmälzle et al., 2013), and have been associated with visual (occipital lobe; Hasson et al., 2004) and auditory processing (lateral temporal lobe; Janata, Tillmann, & Bharucha, 2002). We also found that all PSAs elicited ISC in a caudal, medial region of the frontal lobe. Neural alignment during viewing of natural stimuli is less common in frontal regions, but past studies have observed it in response to emotionally evocative films (Jääskeläinen et al., 2008), stirring political speeches (Schmälzle et al., 2015), and personal narratives (Grall et al., 2021), among others. It has been suggested that ISC in these regions indicates some degree of shared subjective experience (Jääskeläinen et al., 2008) or social and motivational processing (Grall et al., 2021).

We also found that the PSAs that elicited ISC across greater portions of the cortex were rated as more effective by our survey participants. This aligns with the findings of Imhof et al. (2020), who found that effective anti-alcohol PSAs elicited more ISC than ineffective ones. These authors also observed ISC in the occipital lobe and near the temporal-parietal junction for all PSAs but found that this alignment extended to medial regions like the precuneus, PCC, and dmPFC in response to effective messages. They hypothesized that alignment in these regions might index perceptions of self-relevance and risk assessments, indicating whether a message

"got under the skin" (p. 1194) of participants and engaged higher-order cognitive and affective processes. A host of past research supports the interpretation of frontal and medial ISC as an index of higher-order attention and engagement (Dmochowski et al., 2014; Schmälzle et al., 2015; Schmälzle & Grall, 2020). Thus, we propose a similar interpretation of our own results: that participants tended to rate PSAs as more effective when those PSAs contained features that collectively engage audiences' higher-order processes.

Self-reported reactance indicators align with ISC in certain brain regions

In the hopes of identifying neural correlates of reactance processes, we hypothesized that PSAs that elicited higher ratings of self-reported processes would also elicit higher levels of ISC in certain *a priori* regions of interest. Because we hypothesized that these regions were uniquely associated with their respective processes, we also tested the effects of three control ROIs for the purpose of discriminant validity. We expected to see null or negative associations between reactance rank-orderings and rank-orderings of positive valuation regions, because these are thought to index processes that are associated with message success. To support our claim that associations between rank-orderings were due to differences in higher-order cognitive and affective processes, we also tested the visual cortex, where ISC was likely to emerge but was expected to index only visual processing.

Only one set of results aligned with our expectations: a positive association between rank-orderings by anger and those by ISC in the anterior insulae. We also found that rank-orderings by anger were negatively associated with those of ISC in the ventral striatum, which was one of our two positive valuation control regions. This finding aligns with past task-based work associating AI activation with anger and other negative emotions (Alia-Klein et al., 2020; Gilam & Hendler, 2017), and may suggest that messages that elicit more anger are perceived as

less valuable. It also seemingly conflicts with some past ISC work—when Imhof et al. (2017) compared the ISC maps of effective vs. ineffective anti-alcohol PSAs, they found that effective PSAs elicited more ISC in the right AI. This may be explained by the considerable diversity of cognitive and affective experiences that have been associated with the insulae. Far from exclusive to the experience of anger, these regions are thought to be associated with a process of *interoception* by which the brain incorporates information about body states into its conscious state (Craig, 2009). Thus, while the evidence from our study suggests that alignment in the AI can serve as a suitable proxy measure for a message's tendency to elicit anger, this may not be the most appropriate interpretation of this alignment in all cases.

Results did not support our hypothesis that messages that elicited more self-reported negative cognitions would also elicit more ISC within the dlPFC. However, we did observe and inverse association negative cognition rank-orderings and the rank-orderings by both value regions: VS and vmPFC. These regions are often thought to comprise a valuation network in the brain, where inputs from various other brain regions are synthesized into a common value signal that in turn informs decision-making (Scholz et al., 2017). Our findings suggest that messages that elicit more positive cognitions also exert a stronger common influence on these valuation regions, so ISC in these regions during the processing of persuasive messages may index the degree to which people form positive thoughts about those messages and perceive them as valuable.

Another way that our results confounded our expectations was in the observation that a rank-ordering of PSAs by ISC in the visual cortex was associated with rank-orderings by both anger and freedom threat. Some research has observed increased activation in the visual cortex in response to emotional or threatening stimuli (Bradley et al., 2003), so this alignment may have

indicated that participants paid more visual attention to stimuli they perceived to be threatening. The available population of professionally-produced anti-cannabis PSAs did not allow us to control for the visual complexity of our stimuli, however, so this alignment may have been affected by some unrelated element of the messages' visual design.

Self-reported negative cognitions, not neural activity, vary across fear/humor appeals

Contrary to our expectations, we found that most of our measures of reactance processes did not vary significantly across our two appeal conditions. In the case of anger, examination of the model's random effects suggests this may be due to high levels of heterogeneity in anger across both participants and messages. We also observed no significant differences between the average levels of ISC elicited in the AI and dlPFC across messages from our two appeal conditions. This might again be explained by heterogeneity across the messages, or by low power due to the very small sample size (N = 8) for this test.

These findings differ from past results that have found humor appeal messages elicit less reactance than fear appeal messages (Alhabash et al., 2022; Moyer-Gusé, Robinson, & Mcknight, 2019; Sklaski et al., 2009). We suggest they be interpreted with the consideration that absence of evidence does not necessarily imply evidence of absence—we observed considerable heterogeneity across our relatively small sample of fear/humor appeal messages, so it is possible that the effects observed in past studies would also be observed in a larger sample of anti-DUIC PSAs.

Limitations and areas for future research

The results from this study contribute to both the understanding of psychological reactance and audience responses to anti-DUIC messaging, but these contributions must be considered in the light of certain methodological limitations. Most notably, we did not

manipulate the features of any messages for the purpose of instilling or preventing experiences of reactance. Past studies of reactance have often done this by including threatening or dogmatic language (Quick & Stephenson, 2008; Miller et al., 2007), but our procedure presented all messages exactly as they were broadcast. This improved the ecological validity of our findings, but it prevents us from making any inference about the actual causes of variance in either ISC or the value of self-report scales. While observed differences in our variables of interest across stimuli suggest that they were affected by some features of the PSAs' content, we cannot confidently claim any specific feature or set of features as the cause. In order to produce findings that may inform the design of real-world anti-DUIC messages, we suggest that future studies explicitly test message features known to affect reactance such as the use of dogmatic language, fear/humor appeals, and the inclusion of freedom-restoring postscripts.

Second, because this study was intended primarily as a test of message-level relationships, we did not consider the role of individual differences outside the demographic bounds of our sample. Reactance to messages is known to vary across individuals as a factor of trait reactance and personal involvement with the message subject, for example, so consideration of these might reveal important differences in the effects of certain message features. Given that reactance to anti-cannabis messages may partly be a result of pro-cannabis attitudes or distrust or anti-cannabis claims, we suggest future studies seek to identify message which individual differences among audience members are likely to interact with message features to induce or prevent reactance.

Finally, this study hypothesized that the neural correlates of reactance processes would be constrained to specific brain regions. This approach allowed us to select brain regions that had been consistently associated with similar but not identical processes, which was appropriate

given the dearth of past research on this topic. However, it may not align with the reality of reactance's manifestation in the brain. Though certain brain regions are reliably associated with particular cognitive or affective processes, there is no true one-to-one relationship between any one brain structure and process (Genon, Eickhoff, & Kharabian, 2022). Instead, processes are thought to emerge over time from interactions between networks of brain regions (Pessoa, 2014). Given that our findings suggest that the complex processes of psychological reactance can indeed be observed in the brain, we encourage future work use a network neuroscience approach (Fisher et al., 2021) to better understand the spatiotemporal dynamics that underlie these processes.

Conclusions

The purpose of this study was to test hypothesized neural correlates of three processes related to psychological reactance (freedom threat perceptions, anger, and negative cognitions) by comparing responses to a set of anti-DUIC PSAs across an fMRI sample and an independent survey sample. Results suggested that PSAs aligned neural activity in audiovisual, language, and emotional processing regions, and that this alignment was stronger in response to messages that were rated as more effective. Results did not support our hypothesis that freedom threat perceptions could be indexed by activity in the temporal-parietal junction or that negative cognitions could be indexed by activity in the dorsolateral prefrontal cortex but did suggest that messages that align neural activity within the anterior insulae also induce higher levels of anger. In all, these results suggest that observations of shared processes in the brain can be used to identify messages that are likely to induce certain reactance experiences in audiences.

CHAPTER 4 – PICTORIAL WARNING LABELS REDUCE SHARING INTENTIONS, SELF-RELATED PROCESSES ASSOCIATED WITH SOCIAL MEDIA POSTS PROMOTING EDIBLE CANNABIS PRODUCTS

Introduction

The legalization of recreational and medicinal cannabis in Canada and parts of the United States has led to a proliferation of new cannabis products, including an increase in cannabis-infused food/drink products called "edibles" (Borodovsky, 2020). Users perceive edibles as having fewer risks and more benefits than combustible cannabis (Ngyuen et al., 2022), and perceptions of these products' health risks are decreasing as they become more popular with consumers (Reboussin et al., 2019). Use of edibles in place of combustible cannabis may reduce the risks of certain respiratory conditions (Russell et al., 2017), but these products still present risks associated with THC consumption such as impaired driving, onset of mental health conditions, suicidal ideation, and others (National Academies of Sciences, Engineering, and Medicine, 2017). Cannabis edibles also pose a particularly high risk for overconsumption or accidental consumption due to their candy-like appearance and the delayed effects of their active ingredients (Grewal & Loh, 2020). These risks are partly associated with the packaging and marketing materials of cannabis edibles, which are often highly attractive to children (Langrand, Dufayet, & Vodovoar, 2019) and generally present limited content about health risks (Ventresca & Elliott, 2022).

Risks associated with dissemination of cannabis advertisements

Posts promoting cannabis products are considered appealing by adolescents (Liu et al., 2020), and exposure to these posts has been associated with increased intention to use cannabis (Willoughby et al., 2023). Dissemination of these posts on social media has been described as a "critical public health issue" (Moreno, 2022, p. 2), partly because social media platforms allow

users to share posts to their own pages, bypassing any age restrictions imposed on the pages of cannabis vendors (Moreno et al., 2017). Though most social media platforms restrict or ban paid promotions of cannabis products (Berg et al., 2023), manufacturers and vendors alike increasingly promote these products through posts on their public social media pages (Marinello, 2023; Moreno et al., 2018). Policies governing this practice are inconsistent and vague (Berg et al, 2023), allowing cannabis companies a way to reach potential young customers while bypassing state and federal laws about cannabis product advertising. Thus, developing interventions aimed at reducing the dissemination of cannabis advertisements is an important public health concern.

Cannabis warning labels

One simple and relatively low-cost way for regulators to spread awareness about the health risks of legal products like cannabis edibles and reduce sharing online is by requiring warning labels on packaging and advertisements. Though states that allow the sale of cannabis typically mandate the presence of health warnings on both products and advertisements (Kruger et al., 2022), content analyses have found that social media posts promoting cannabis products usually do not contain such information (Moreno et al., 2017, 2022).

Little is known about the effects of warning labels on engagement with posts promoting cannabis products, but research in other domains suggests labels can have meaningful effects on engagement and attitudes. For example, when Phua, Lin & Lim (2018) added warning labels to celebrity endorsements of e-cigarettes on Instagram, they found the presence of warnings was associated with less favorable attitudes toward the ads and with reduced intentions to use e-cigarettes. Similarly, Wu et al. (2023) found that adding warning labels to influencers' Instagram posts promoting cigarillos was associated with lower levels of engagement. Regulations

governing the presence of warning labels on such posts are currently unclear (Wu et al., 2023), and the U.S. National Cancer Institute has publicly called for more research into the effects of pictorial warning labels on sharing of social media posts about tobacco products (Thrasher et al., 2019). Thus, the impacts of health warning information on engagement with social media posts is considered an important but understudied public health concern, partly because this information can affect the ways people share and otherwise engage with social media content.

The current study aimed to examine the impact pictorial cannabis warning labels (CWLs) have on cannabis advertisement processing and whether CWLs reduce engagement and online sharing among at risk youth. Specifically, this study had three aims: 1) to test whether pictorial CWLs reduce intentions to engage with and share marketing materials for edibles on social media; 2) to test whether CWLs influence key brain regions during message exposure associated with sharing; and 3) to conduct an exploratory analysis to examine how CWLs influence the underlying neural processes associated with cannabis advertisement processing.

Sharing and self, value, and social processing brain networks

Recent research using neuroimaging suggests that the decision to share information online can be explained by a neural model of value-based virality, in which information about the message, reflections on the self, and inferences about likely responses from others are incorporated into a single, domain-general signal of value (Scholz et al., 2020; 2017). The strength of this signal is thought to in turn predict the decision to share or not to share a particular piece of information or content. For example, previous research found that increased activity in self, social, and value-processing networks when encoding real-world news stories was associated with an increased likelihood to share with others in a sample of fMRI participants (Baek et al., 2017), this activity also predicted which news stories were shared online more often

at the population level (Scholz et al., 2017). We review each of those networks below, explaining their connection to their respective processes and the specific role they are hypothesized to play in the sharing process.

Self processing

In their neural model of valuation and virality, Scholz et al. (2020) associated self-related processing with activity in a portion of the medial prefrontal cortex (mPFC) and a cluster overlapping the posterior cingulate cortex (PCC) and precuneus. This aligns with results of meta-analyses seeking neural correlates of self-reflection and trait attribution (Murray et al., 2012; Northoff et al., 2006), which found that these regions were reliably activated when participants made judgements about whether certain traits were applicable to themselves. The connection between the mPFC and self processing has since been supported by hundreds of subsequent studies (e.g. Lieberman et al., 2019; D'Argembau, 2013), and this region has been specifically implicated in self-related processing during the encoding of media messages (Scholz & Falk, 2018). Activity in the mPFC in self-processing contexts is often accompanied activity in the posterior cingulate (PCC) and precuneus (Johnson et al., 2002; Northoff et al., 2006), which have been consistently associated with self-related processes like autobiographical memory and first-person perspective taking (Cavanna & Trimble, 2006).

In outlining their neural model of value-based virality, Scholz et al. (2020) suggest that self-processing regions engage before the decision to share because participants are considering the likely benefits to their social status and esteem that come with sharing this information. This interpretation is consistent with behavioral findings that people are often motivated to share information online by a desire to enhance their self-presentation (Barasch & Berger, 2014). Interpreting a similar set of findings, Baek et al. (2017) ascribed activity in self-processing

regions to perceptions of personal value and self-relevance, noting that these regions predicted not only the decision to share articles but also the decision to select articles for oneself.

Social processing

To measure social-cognitive processes, the neural model of value-based virality (Scholz et al., 2020) captures activity across a wide range of regions in the frontal, temporal, and parietal lobes of the brain. Specifically, the network includes a cluster stretching from ventral to dorsal portions of the mPFC, the bilateral temporal-parietal junction (TPJ), and the precuneus This network was derived from results of a past study that tested the neural correlates of a "theory of mind" task in a large sample of neurotypical and autistic participants (Dufour et al., 2013). The network largely overlaps with several other prominent findings for correlates of theory of mind and other social cognition tasks (e.g. Yang et al., 2015), and encompasses almost all regions associated with mentalizing in a recent meta-analysis of 105 fMRI studies (Arioli et al., 2021). In other words, the regions included in this network are reliably associated with the process of making inferences about the thoughts and feelings of others.

Activity in these social processing regions before sharing decisions is thought to indicate the degree to which people are considering the utility of the message to relevant others and anticipating their likely responses (Scholz et al., 2020a). This interpretation is supported by a similar analysis performed by a similar team of authors (Scholz et al., 2020b), which found that activity in social processing regions was stronger when people were considering narrowcasting a message (i.e. sharing it directly with close friends) than broadcasting (sharing to a social media page). Past work has also found that participants were more likely to spread ideas that elicited stronger activation in social processing regions and that people who showed more activation in these regions during the encoding of ideas were more effective at spreading those ideas in a post-

scan exercise (Falk et al., 2013). Activity in these regions has also been shown to predict intentions to share marketing videos on Facebook (Motoki et al., 2020).

Value processing

The value processing network described in Scholz et al.'s (2020) neural model of value-based virality includes the ventromedial prefrontal cortex (vmPFC) and the ventral striatum (VS). These regions have been associated with valuation processes in hundreds of neuroimaging studies (Bartra et al., 2013), and activation in these regions is thought to indicate a domain-general value signal that guides decision making (Levy & Glimcher, 2012). Activity in the vmPFC in particular is thought to indicate assignments of subjective value to the content of media messages (Scholz & Falk, 2018), and this activation has been shown to predict the success of persuasive messages at both the individual (Falk et al., 2010) and population (Berkman, Falk, & Lieberman, 2012) levels. As shown in Chapter 2 of this dissertation, activation of this region also tracks with participants' ratings of experimental stimuli.

In the context of sharing, Scholz et al. (2020; 2017) suggest that activation of value processing regions indicates the integration of information from self- and social- processing regions into a coherent, domain-general value signal that is then used to inform the sharing decision. While considerations like self-relevance and possible effects to one's reputation or relationships involve value judgments that are non-quantitative, these authors suggest that the brain's value system converts these judgments into a common currency—much as a currency exchange might convert both Colombian Pesos and Swiss Francs into U.S. Dollars. The value of that currency is thought to track with activity in the value network, and it can be used to "purchase" a sharing decision in the sense that greater activation of this network increases the

likelihood of sharing. In short, the value system is thought to mediate the effects of self and social processing on the decision to share a piece of media or information.

Effects of pictorial warning labels on brain activity

Though the effects of warning labels on sharing behaviors has not been studied using neuroimaging methods, these methods have been used to better understand the effects of warning labels more generally. Research into the effects of graphic warning labels designed for use on tobacco products suggests these labels can drive activity in both cognitive and affective brain regions. For example, Wang et al. (2015) found that warning labels that had previously been rated as emotion-inducing elicited activity in the hippocampi, amygdala, and insulae—regions that have been associated with memory (Eichenbaum et al., 1999) and affective processing (Barrett & Bliss-Moreau. 2009). The content of these labels was also better remembered by participants, and exposure to them was associated with weaker urges to smoke. Several other studies have observed activation of affective brain regions during the encoding of graphic warning labels by smokers (Green et al., 2016; Riddle et al., 2016; Owens et al., 2016; Do & Galván, 2015). Some warning labels have also been shown to elicit activity in the vmPFC, and activity in this region in response to warning labels has been shown to predict smoking cessation after the scanning session (Owens et al. 2016; Riddle et al., 2016). These latter findings are consistent with a growing body of research that suggests vmPFC activation during message encoding might index perceptions of self-relevance and/or subjective value that inform behavior (Falk & Scholz, 2018). Notably, this region also overlaps with portions of the self- and value-processing networks thought to be responsible for message virality (Scholz et al., 2017; 2020).

Importantly, these studies either displayed warning labels in isolation (Wang et al., 2016; Riddle et al., 2016; Do & Galván, 2015) or paired them with images of cigarette packaging

(Owens et al., 2016; Green et al., 2016). Thus, little is known about the ways these labels might impact the encoding of messages that have also been designed to elicit cognitive and affective responses, such as marketing messages. Theorizing that associates activation of self-, social-, and value-processing networks with an integrated, domain-general value signal (Scholz et al., 2020) might suggest that the addition of negatively-valenced stimuli like warning labels would downgrade the signal elicited by positively-valenced marketing messages, but this possibility has not been explicitly tested.

Objective, aims, and hypotheses

The objective of the current study is to triangulate whether the addition of pictorial CWLs to online marketing messages promoting edible cannabis products might affect audiences' sharing behaviors and the patterns of neural activity associated with those behaviors. Online dissemination of cannabis marketing messages is recognized as an emerging public health threat (Moreno, 2020), and achieving a better understanding of the processes that underlie that dissemination can inform an important and ongoing public policy discussion. Further, this work sought to build on recent work that explored the interaction between social media sharing and processes that can be observed in the brain. Toward these ends, this study had three specific aims:

Aim 1: To test the effects of CWLs on broadcasting and narrowcasting intentions

The first aim of this study was to test whether test whether the addition of pictorial CWLs to social media posts promoting cannabis products might reduce peoples' intentions to share those posts. The ability of users to spread this pro-cannabis content to friends or to their larger networks has been described as a critical risk to public health (Moreno, 2022), and adding

warning labels to similar posts has been shown to reduce social media engagement (Phua, Lin & Lim, 2018).

H1a: Participants exposed to posts that are paired with pictorial warning labels will report lower intentions to share posts with close friends (narrowcasting).

H1b: Participants exposed to posts that are paired with pictorial warning labels will report lower intentions to share posts with their broader social networks (broadcasting).

H1c: Participants exposed to posts that are paired with pictorial CWLs will report lower intentions to engage in other virality behaviors (liking and replying)

Aim 2: Test whether the presence of CWLs affects neural networks relevant to social media processes

The second aim of this study was to examine whether CWLs can reduce activity in regions previously associated with increased sharing during the encoding of online cannabis marketing messages. Past research has found that messages that are more likely to be shared online also tend to elicit increased activity in networks associated with self, social, and value processing (Baek et al., 2017; Scholz et al., 2017).

H2a: Participants undergoing fMRI will exhibit less activity in the self-processing network while encoding messages that are paired with CWLs than while encoding messages that are not paired with CWLs.

H2b: Participants undergoing fMRI will exhibit less activity in the social-processing network while encoding messages that are paired with CWLs than while encoding messages that are not paired with CWLs.

H2c: Participants undergoing fMRI will exhibit less activity in the value-processing network while encoding messages that are paired with CWLs than while encoding messages that are not paired with CWLs.

Aim 3: Identify neural correlates of the effects of CWLs on encoding of marketing messages

Past research has associated the encoding of graphic warning labels with activation of both cognitive and affective brain regions (Wang et al., 2015; Owens et al., 2016; Green et al., 2016; Riddle et al., 2016; Do & Galván, 2015), but little is known about the ways these warning labels might affect the encoding of positively-valenced marketing messages. Thus, we sought to identify brain regions that show sensitivity to the presence of warning labels during the encoding of social media cannabis marketing posts.

> *RQ1: What regions of the brain show differences in activation when cannabis marketing messages are presented alongside graphic CWLs compared to messages that do not contain CWLs?*

Methods - Study 1

Participants

In summer 2022, 1,776 participants between 18 and 25 (M = 21.56, SD = 2.11) were recruited through an online Qualtrics panel. All participants responded positively to at least one of the following pro-cannabis questions: "Would you try marijuana if one of your best friends offered it to you?", "Do you think you would use marijuana in the next 6 months?", and "Are you curious about using marijuana?". Most participants identified as White (50.39%), followed by Black (30.57%), or another race (19.03%), and were mostly women (56.93%).

Survey protocol

After consenting, participants completed a set of demographic questions and were then randomized into one of ten conditions in this 3 (No CWL, Text-only CWL, Pictorial CWL) x 3 (No comments, pro-cannabis comments, anti-cannabis comments) + 1 (no-message control) factorial design[2]. In each condition, participants saw a set of three cannabis marketing posts

[2] To maintain consistency with Study 2, only No CWL and Pictorial CWL conditions were analyzed for this chapter

drawn from a pool of 90 real-world marketing messages that was either presented alone of presented with a CWL. After exposure to each stimulus, participants were asked to report their intentions to engage in viral behaviors. This study was reviewed and approved by the university Institutional Review Board of the corresponding author.

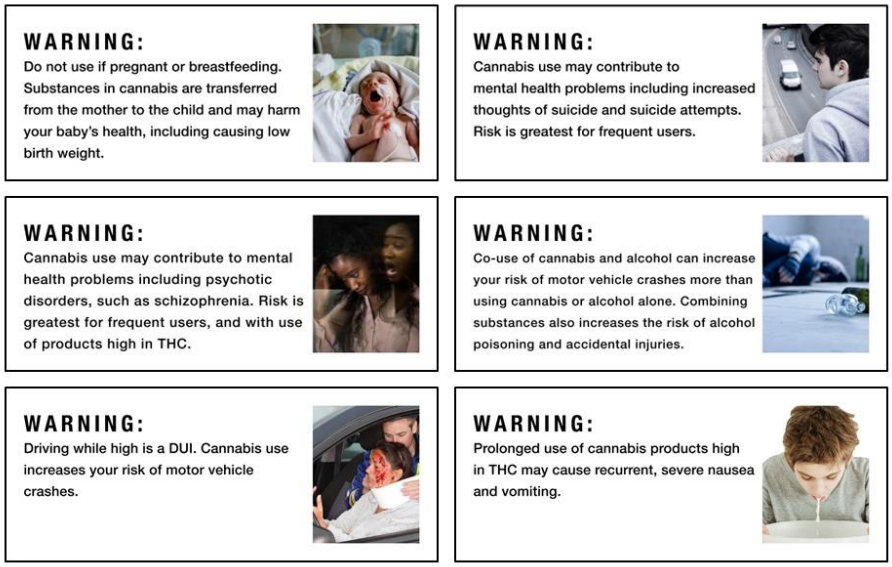
Figure 4.1: Stimuli presentation

Note: *An example of cannabis marking posts as presented in the No CWL (left) and Pictorial CWL (right) conditions*

Pictorial CWL content

Pictorial CWLs were designed to describe established health risks (National Academies of Sciences, Engineering, and Medicine, 2017) specifically suggested for presentation on CWLs in the state of California. These risks include: cognitive function loss tied to early use, driving risks, mental health issues, suicidal ideation, exposure to toxic contaminants, nausea and vomiting, fetal transfer of substances, and risks associated with the delayed onset of psychoactive effects of cannabis edibles:

Figure 4.2: Pictorial CWLs



Note: *Six of the 10 pictorial CWLs used in studies 1 and 2.*

Viral behavior intentions

Participants reported viral behaviors for each of the three marketing posts by responding to the following prompts one a Likert-style scale of 1 (not at all likely) to 7 (extremely likely):

1. …share this post directly with someone you know (via email, direct message, etc.) (narrowcasting)
2. …share this post to your own social media timeline (e.g., Facebook, Twitter) so that many people could see it? (broadcasting)
3. …click the "like" button? (liking)
4. …reply to this post? (replying)

Methods - Study 2

Participants

A total of 40 participants were recruited for this study using the job board of a large midwestern research university. All participants were between 18 and 24 years of age (M = 20.1, SD = 1.34) and scored at least a 2 on the Cannabis Use Disorder Identification Test (CUDIT). Participants scored an average of 24.38 on this measure (SD = 7.78), and 100% exceeded the

threshold for CUD. The sample was predominantly white (71.79%) and female (50.00%). All participants were right-handed, had normal (or corrected to normal) vision, were not taking any psychoactive medications, did not suffer from claustrophobia, and did not have metal in their bodies that was contraindicated for MRI.

Table 4.1: Sample demographics

Study	*N*	*Age*	*Sex*	*Race*	*Methods*	*Outcomes*	*Population*
Study 1	1,776	21.56 (2.11) 18-25	Female 56.93% (n = 1,011)	White (50.39%) Black (30.57%) Another race (19.03%)	Online-survey experiment	Narrowcasting Broadcasting Liking Posts Replying to Posts	At risk for cannabis use.
Study 2	40	20.10 (1.34) 18-24	Female 50.00% (n = 20)	White (71.79%) Black (7.5%) Another race (27.5%)	fMRI	ROI Whole Brain Analysis	Mean cannabis use measure

fMRI protocol

While undergoing two consecutive eight-minute runs of fMRI, participants viewed 60 cannabis marketing posts drawn pseudo-randomly from the same sample used for Study 1[3]. Each post was paired with one of nine pictorial warning labels (described in the following section) or with no warning label using a gray rectangle of equal dimensions. Post/label pairings were assigned using a randomized, counterbalanced design. Participants viewed each post/label pairing for a period of seven seconds, then viewed a set either a set of social media comments that were manipulated to present positive or negative attitudes toward cannabis use, or a visually

[3] Four participants only underwent one run due to technical errors, viewing a total of 30 post/label parings each.

similar set of comment boxes containing lorem ipsum filler text. Participants viewed these comments for a period of six seconds, then responded to the prompt “How effective is this cannabis ad?” on a scale of 1 (not effective) to 4 (very effective) using a four-button Current Designs response pad held in the participant’s right hand. Participants were given three seconds to respond to this prompt, then were presented with a fixation cross for a randomized, jittered period of an average of 1.5 seconds.

fMRI data acquisition

Structural and functional brain imaging was conducted using a 3 Tesla GE Discovery MR750 scanner. Head motion was minimized with foam padding on the head coil. The structural imaging procedure was changed mid-study to reduce scan time. For 3 participants, structural scans were obtained using a motion-corrected T1-weighted MPnRAGE acquisition with 1.0 mm isotropic spatial resolution (Kecskementi et al., 2018). For the remaining 37 participants, scans were obtained using a FSPGR BRAVO sequence. Two functional runs were recorded (TR = 800ms, TE = 20ms, flip angle = 60°, matrix size = 96x96, 54 axial slices, 3mm thick; voxel size = 3.0x2.5x2.5) for 36 participants, and one run was recorded for the remaining four participants. Stimuli were projected onto a screen on the back of the scanner bore, which participants viewed through a mirror mounted directly above the head coil. Participants heard sound through MR-compatible headphones designed to operate within the bore.

Preprocessing was performed using the afni_proc.py program within the Analysis for Functional Neuroimaging (AFNI) software package (Cox, 1996). In this process, slice timing correction was performed using the first slice as reference, and images were corrected for minor motion by spatially realigning all images to the first acquired volume. Structural images were registered to the main functional image and skull-stripped. Functional and structural runs were

warped to align with the Montreal Neurological Institute MNI151 template brain and smoothed with a 4mm Gaussian kernel.

Data were modeled at the single subject level using the general linear model as implemented in AFNI. Two trial types were modeled during the 7 second exposure to the cannabis ads (cannabis ads with CWLs, cannabis ads without CWLs). In addition, we modeled exposure to the jittered fixation cross (i.e., rest). In total, three trial types were modeled (cannabis ads with CWLs, cannabis ads without CWLs, and rest), as well as random effects, motion and nuisance regressors.

A priori networks of interest

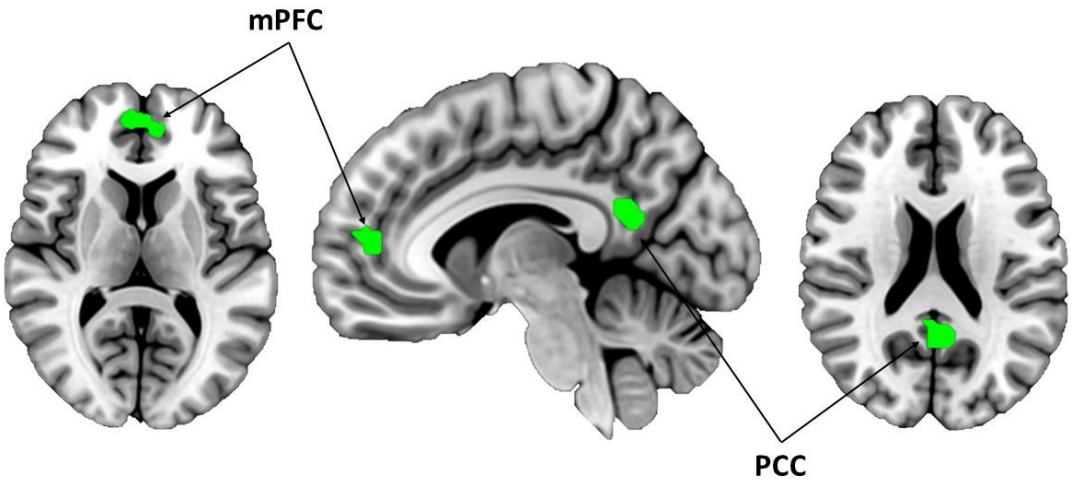
The self, social, and value ROIs for the current study were based on Scholz et al. (2017) in their initial test of the model of value-based virality and were obtained from that research team. Each network included clusters for each of the regions described in the above literature review. Because parts of the social processing network overlapped with frontal portions of both the self and value networks, voxels that were part of these networks were removed from the social network mask.

Table 4.2: Network of interest coordinates

	X	*Y*	*Z*	*Size (voxels)*
Self				
mPFC	-7.5	-53.2	-2.0	164
PCC	0.0	59.2	13.0	144
Social				
TPJ	*-47.5* *62.5*	-3.2 64.2	-47.0 8.0	3846 1700
PFC	10.0	-43.2	20.5	2424
Precuneus	5.0	61.8	10.5	1894
Value				
vmPFC	-10.0	-8.2	-12.0	235
VS	2.5	-35.8	-17.0	233

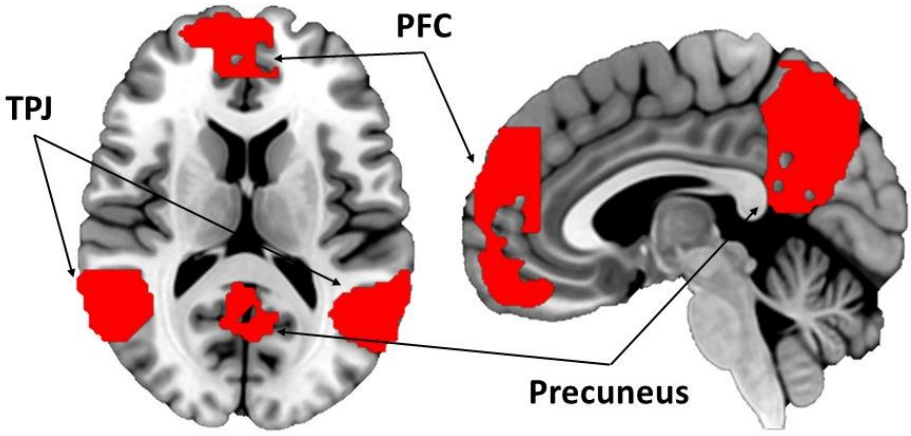
Note: *X, Y, and Z coordinates are presented in the MNI template brain*

Figure 4.3: Self processing network



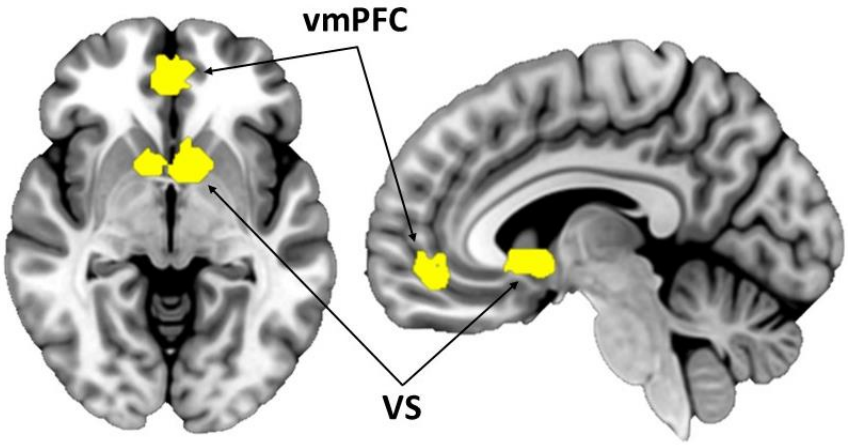
Note: *The self processing network is made up of clusters in the vmPFC and PCC (Baek et al., 2017)*

Figure 4.4: Social processing network



Note: *The social processing network is made up of large clusters in the PFC, TPJ, and precuneus. Voxels that overlap with self and processing networks were removed (Baek et al., 2017)*

Figure 4.5: Value processing network



Note: *The value processing network is made up of two clusters: one in the vmPFC and another in the VS (Baek et al., 2017).*

Analyses

Aim 1: To test the effects of CWLs on viral engagement behaviors

To test whether intentions to perform sharing and other viral behaviors was affected by the presence of pictorial CWLs, we conducted tested the fit of a series of linear mixed effects models in which behavioral intentions served as the outcome variable and warning label

condition served as the predictor variable. To account for the clustered nature of our data, we also modeled by-participant and by-marketing post random intercepts and slopes.

Aim 2: Test whether the presence of warning labels affects activity in a priori networks of interest

To test whether the presence of warning labels affected the degree to which participants engaged in self-relevance, social, and value processing, we tested the fit of three multilevel linear regression models. In each model, the activity within the network of interest during encoding of ads (compared to a fixation cross baseline) served as the outcome variable and warning label condition (present/absent) served as the predictor variable. To account for the clustered nature of our data, we included by-participant and by-stimuli random slopes and intercepts for warning label condition.

Aim 3: Identify brain regions that are sensitive to the presence or absence of warning labels

To identify brain regions outside or *a priori* networks of interest that might be sensitive to our warning label condition, we conducted an exploratory whole-brain analysis using the AFNI command 3dttest++ that contrasted the encoding of posts that were paired with warning labels compared to posts that were paired with no warning labels (i.e., gray rectangle). The result of this analysis was a whole-brain map in which each voxel contained a value representing the effect of our warning label condition on the neural signal in that voxel. All whole brain analyses are reported with a threshold of ($p = .005$, $K > 46$), corrected for multiple comparisons based on Monte-Carlo-style simulation (10,000 iterations) using the AFNI command ClustSim, corresponding to a corrected threshold of $p < .05$.

Study 1 Results

Presence of pictorial CWLs is associated with lower intentions for viral behaviors

The presence of pictorial CWLs on social media posts promoting cannabis products was associated with lower intentions to "like" the post β = -0.160, $F(1, 1234.4) = 8.742$, $p = 0.002$, to share the post with specific friends (narrowcasting) β = -0.183, $F(1, 1234.6) = 14.255$, $p < 0.001$, and to share the post with a larger social media network (broadcasting) β = -0.1.60, $F(1, 1234.4) = 8.742$, $p = 0.002$. Complete results are illustrated in Figure 4.6 and presented in Table 4.3 below.

Figure 4.6: Effects of pictorial CWLs on intentions to perform virality behaviors

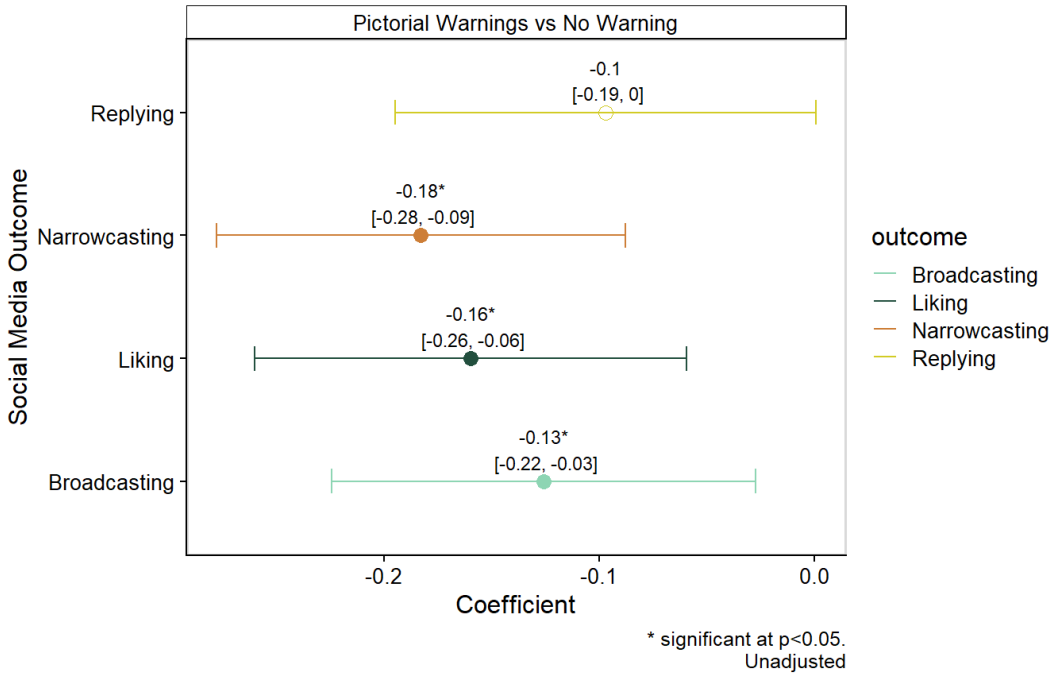


Table 4.3: Effects of pictorial CWLs on intentions to perform virality behaviors

	$B_{CWL > No\ CWL}$	*F*	*df*	df_{resid}	*p*
Replying	-0.097	3.797	1	1234.9	0.052
Narrowcasting	-0.183***	14.255	1	1234.6	< 0.001
Liking	-0.160**	8.742	1	1234.4	0.002
Broadcasting	-0.126*	6.276	1	1234.5	0.013

Study 2 Results

Presence of CWLs is associated with less activity in the self processing network

After testing the effects of warning label condition on activation of our networks of interest, we found that participants exhibited significantly decreased activity in the self processing network while encoding messages that were paired with warning labels β = -0.085, $F(1, 24.567) = 5.497$, $p = 0.027$. Warning labels did not show any significant effects on activity in social or value processing networks.

Table 4.4: Effects of pictorial CWLs on network activation

	$B_{LABEL > NO}$	*F*	*DF*	DF_{RESID}	*P*
SELF	-0.085*	5.497	1	24.567	0.027
SOCIAL[1]	-0.026	1.742	1	1629.1	0.1871
VALUE	-0.071	1.780	1	24.682	0.194

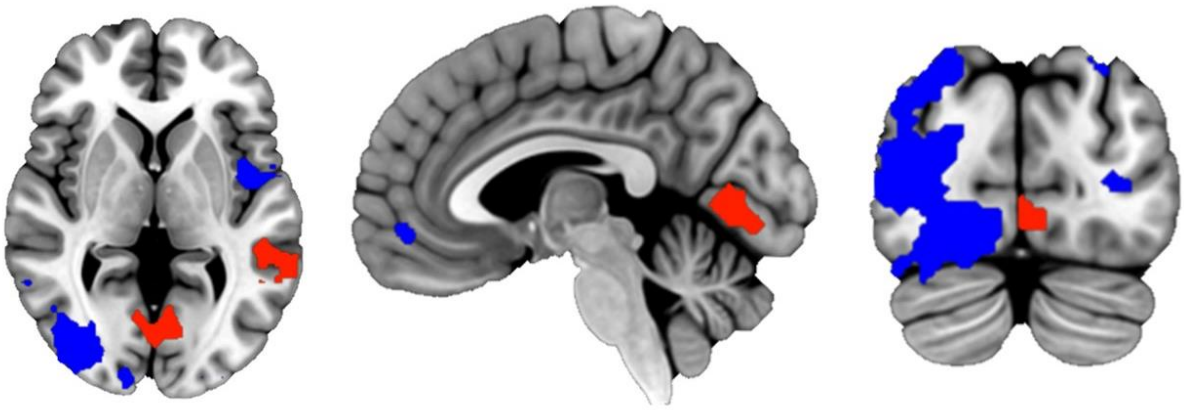
[1] *The initial multilevel model for this outcome failed to converge. This was addressed by following the procedure recommended by Brauer & Curtin (2018),ultimately removing by-cue and by-subject random intercepts.*

Presence of pictorial CWLs affects signal in visual, language, and socio-emotional processing regions

Exploratory whole-brain analysis revealed several clusters in the brain where neural activity was affected by the presence or absence of warning labels (warning labels > no warning labels). Results indicated that there was significant increased activation in the medial ventral proportion of the occipital lobe and in the left TPJ during messages that contained a graphic

CWL compared to messages that did not. In addition, results indicated that there was significant decreased activity in the left superior temporal gyrus, dorsolateral prefrontal cortex (dlPFC), mPFC, and posterior portions of the visual cortex.

Figure 4.7: Clusters affected by Pictorial CWL > No CWL contrast



Note: *Illustrative axial (left), sagittal (middle), and coronal (right) slices from the whole-brain analysis contrasting Pictorial CWL > No CWL conditions. Red clusters indicate higher activation during encoding of messages with CWLs, blue clusters indicate lower activation.*

Table 4.5: Neural activity associated with exposure to cannabis warning labels.

REGION	X	Y	Z	Z score	voxels
occipital lobe	-40	81.8	-19.5	-4.342	2537
middle occipital lobe	32.5	94.2	18	-4.311	480
inferior parietal lobe	52.5	44.2	58	-3.396	191
superior parietal lobe	-45	44.2	63	-3.319	146
inferior temporal gyrus	60	64.2	-17	-3.249	110
insula, superior temporal gyrus	50	1.8	0.5	-3.991	104
dorsal medial prefrontal cortex	-17.5	-45.8	50.5	-3.622	71
superior frontal gyrus	25	-20.8	63	-3.201	67
superior temporal gyrus	65	1.8	-4.5	4.051	61
ventral medial prefrontal cortex	2.5	-53.2	-9.5	-2.889	54
inferior frontal gyrus	-50	-43.2	10.5	-2.987	51
fusiform gyrus	30	54.2	-17	-3.554	47
inferior frontal gyrus	-55	-13.2	5.5	-2.901	47
middle temporal gyrus	67.5	51.8	10.5	4.237	287
lingual gyrus	0	71.8	5.5	4.544	239
superior temporal gyrus	65	1.8	-4.5	4.051	61
fusiform gyrus	42.5	44.2	-24.5	2.878	55

Note: *Significant neural activity associated with exposure to cannabis ads with warning labels compared to cannabis ads with no warning labels. Whole brain results are corrected for multiple comparisons using a threshold of ($p = .005$, $K > 46$), corresponding to a corrected threshold of $p < .05$.*

Discussion

Posts promoting cannabis products are considered appealing by adolescents (Liu et al., 2020), and exposure to these posts has been associated with increased intention to use cannabis (Willoughby et al., 2023). Dissemination of these posts on social media has been described as a "critical public health issue" (Moreno, 2022, p. 2) and the U.S. National Cancer Institute has publicly called for more research into the effects of pictorial warning labels on sharing of social media posts (Thrasher et al., 2019). Therefore, the current study aimed to examine whether and how pictorial warning labels might slow the dissemination of social media posts promoting

cannabis products. Using a set of CWLs recently proposed for use in California, we tested the effects of warning label placement on both behavioral sharing intentions and on the activation of neural networks thought to be associated with the decision to share content online. Our results show that the addition of pictorial CWLs reduces behavioral intentions to share product posts and suggest that this may occur because viewers see these posts as less self-relevant or less in line with their conceptualizations of self.

This study contributes to the understanding of an emerging public health issue and to current theoretical discussions about the role of the brain in the processing of media messages. Specifically, these results make three distinct contributions. First, they show that a set of pictorial CWLs recently proposed for use in the state of California can reduce the virality of social media posts promoting cannabis products. Second, they suggest that the presentation of CWLs during message encoding of cannabis advertisements reduces activity in regions associated with self-processing. Third, they show that the presence of CWLs affects a number of brain regions during the encoding of cannabis marketing messages, including areas associated with cognitive and affective processes.

Pictorial CWLs reduce self-reported sharing intentions

In keeping with our expectations, we found that participants who viewed marketing posts that were paired with pictorial CWLs reported lower intentions to share and engage with those post than participants who viewed marketing posts that were not paired with CWLs. This was true of both sharing with close friends (narrowcasting) and sharing with general networks (broadcasting). We also found that participants said they were less likely to “like” posts with warning labels, and participants reported being less likely to comment on or reply to these posts to a degree that approached but did not meet our threshold for statistical significance.

These findings align with past literature suggesting that the addition of warning labels or other health risk of information can reduce engagement with social media content (Wu et al., 2023) and reduce the efficacy of marketing materials (Neiderdeppe et al., 2018). They also have important policy implications, because the viral spread of cannabis marketing content on social media is thought to be a critical emerging public health concern (Moreno, 2020). Marketing materials promoting edibles often feature sweet or candy-like images, which are likely to appeal to children (Langrand, Dufayet, & Vodovoar, 2019). Given that social media sharing allows marketing content to bypass age-restricting filters (Moreno et al., 2017), the viral spread of these messages may facilitate distribution to adolescents and other social media users under the legal age for cannabis use.

These findings also confirm that the presence of warning labels has a direct impact on social media sharing decisions, supporting our use of a CWL/no CWL manipulation in our tests of the neural correlates of these decisions.

Pictorial CWLs are associated with decreased activity in self-processing brain regions

We found that the presence of pictorial CWLs during message encoding was associated with decreased activity in a network of brain regions associated with self processing. Given that the presence of CWLs was also associated with reduced sharing intentions in our survey sample, this finding aligns with past work that has found content that elicits more activity in self processing regions are also more likely to be shared online (Baek et al., 2017; Scholz et al., 2017). These findings were not consistent with Scholz et al.'s (2020a) neural model of value-based morality, however, because we did not observe any differences in activation of the social or value processing networks. This is important because the value-based virality model does not assert that sharing decisions are influenced by activity in these three networks equally. Instead,

the model claims that sharing decisions are directly informed by a domain-general value signal that reflects information from self and social processing regions and is indexed by activity in the vmPFC and VS (Scholz et al., 2017; 2020). Though we did find evidence that CWLs affected one of the two antecedent processes, we did not find that CWLs exerted any effect on activity in the value network. Our findings are also inconsistent with those of other studies of sharing. Falk et al. (2013) found that sharing of information after a scanner session could be predicted by activity in the social and value networks during encoding, and Motoki et al. (2020) found that only activity in the social processing network during the encoding of videos predicted their real-world virality on Facebook.

Our findings differ in some important ways from those of past studies on the neuroscience of sharing, but they align with the thinking that persuasive media messages succeed when they activate processes of self-relevance, self-reflection, and subjective value (Scholz & Falk, 2018). Dozens of studies have associated persuasive success with activation of self processing networks during message encoding, with activation in these regions predicting perceptions of message effectiveness (Chapter 2) and message-consistent behaviors at both the individual (Falk et al., 2010) and population (Falk et al., 2015; 2012) levels. Some have proposed that activity in these regions indexes processes of *subjective valuation* (Scholz & Falk, 2018) or *self-value integration* (Vezich et al., 2018), in which participants recognize message content as having specific value to themselves or their perceptions of their self-concept. Perceptions of self-relevance, as indexed by activation of self-processing regions, have also been proposed as a direct predictor of sharing intentions (Baek et al., 2017). Given that the addition of health warning information has been shown to negatively impact the success of marketing messages (Niederdeppe et al., 2018), it is possible that CWLs' effects on both sharing intentions and self-

processing activity can be explained by a process in which the presence of these labels distracts from or negatively impacts perceptions of the messages' content as self-relevant.

Participants in both samples were screened for likely cannabis use (survey sample) or symptoms of cannabis use disorder (fMRI sample), and all participants in the fMRI sample reported CUDIT values that exceeded the established threshold for cannabis use disorder. Thus, this sample of habitual cannabis users may have been particularly likely to view cannabis marketing materials as self-relevant because they have experience with or would consider buying the products featured. Pictorial warning labels have previously been shown to reduce product appeal (Phau & Lime, 2022) and cravings (Wang et al., 2015) among product users, and our results suggest this may occur because the information presented in these labels disrupt these reflections.

Differences between the findings of this study and those of other studies of sharing might be partly explained by study design. The data used to develop the neural model of value-based virality was obtained using a procedure that sometimes explicitly prompted participants to consider the act of sharing stimuli, either with close friends or with a larger online audience (Baek et al., 2017; Scholz et al., 2017). Thus, participants were primed to consider the act of sharing while encoding stimuli, and this structure may have made participants more likely to consider the relative benefits and potential social consequences of their sharing decisions. Our own procedure contained no such prompt—participants were asked only to consider the marketing messages as they were presented, then to offer a simple rating of their perceived effectiveness. While this procedure arguably better resembles the passive nature of real-world social media scrolling, it also precludes us from describing the neural processes we observed as part of a social media sharing decision. For this reason, we do not consider the findings of this

study to be contrasting with past work. Instead, our findings suggest that warning labels impact self-related processing even during unstructured encoding, which generally precedes the real-world decision to share social media content. This may be an early step in the process of informing sharing decisions, with social and value processes being engaged closer to the moment of the decision itself.

Exploratory whole brain analysis

Finally, we conducted an exploratory whole-brain analysis to determine whether activity in regions outside our *a priori* networks of interest were affected by the presence of CWLs during encoding of cannabis marketing posts. Results found the presence of CWLs was associated with decreased activity in the dmPFC, vmPFC, and inferior frontal gyrus (IFG), among other regions. Some of these regions overlap with our networks of interest or are associated with similar processes. For example, decreased activation in the dmPFC and vmPFC way indicate reductions in social cognition (dmPFC) or positive valuation (vmPFC) processes during encoding. These findings largely align with our hypotheses that presence of CWLs would be associated with decreased activity in social and value processing networks and may suggest that these effects did occur but manifested in a way that was not captured by our approach to measurement.

Activity in the IFG has been associated with inhibitory and attentional control (Hampshire et al., 2010), so our observation of deactivation in this region may suggest participants were not able to inhibit the warning label information when viewing the cannabis ads. Results also indicated that decreased activity in the visual cortex was associated with exposure to cannabis warning labels. Reduced activation of primary visual processing regions may reflect a blunting effect of warning labels on motivated attention to the content of marketing

messages (Bradley et al., 2003). However, these conclusions are just a set of plausible explanations given these regions have a many to one mapping in the brain and may be involved in other processes. Thus, future research should more explicitly test whether these potential neural networks are associated with exposure to warning labels.

We did not observe differences in activity in some regions that past neuroimaging studies have associated with the encoding of graphic warning labels. Several of these studies found that graphic warning labels elicited increased activation in the amygdala (Wang et al, 2015; Owens et al., 2016; Greene et al., 2016; Riddle et al., 2016) and/or insulae (Do & Galván, 2015), suggesting that the presence of warning labels was associated with some form of affective processing. Our own results did not identify any affective brain regions that showed increased activation during the encoding of graphic CWLs. This may partly be a factor of label content—past research has used warning labels designed for use with tobacco products, which often contain images that are especially frightening or disgusting.

Limitations and recommendations for future research

The findings from this study provide interesting and useful insights that can inform both public policy governing the style and use of CWLs and the emerging understanding of the way labels and media messages interact with the brain. However, these results are also limited in some important ways. Most notably, our analyses did not include any tests directly relating activity in our *a priori* brain networks with sharing intentions. Although we observed that the presence of graphic CWLs on cannabis marketing messages was associated with both lower likelihood of sharing and less activation of brain regions associated with self-processing, our methods were unable to provide any direct evidence connecting these two outcomes. Past

research has found support for such connections using a brain-as-predictor framework (Scholz et al., 2017), and we suggest that future research follow a similar approach.

Second, the results of this study are complicated by our choice to test the effects of both graphic CWLs and promotional messages simultaneously. This approach reflects the real-world use of product and advertisement warning labels, which are designed to interact with product packaging or marketing materials, but it allows for the possibility of stimulus confounding. Our design did assess the effects of marketing messages that were not paired with warning labels, but we did not present participants in either the survey sample or the fMRI sample with warning labels in isolation. Given that pictorial warning labels have been shown to affect brain regions that are involved in our self, social, and value-processing networks (Riddle et al., 2016; Do & Galván, 2015), their direct effects on these networks may have interfered with the interaction effects we sought to measure. To address this, we encourage future work to use alternative experimental designs that would allow for easier analysis of the warning labels' isolated effects.

Finally, this study's design did not control for differences in the amount of information presented across the Pictorial CWL / No CWL conditions. Participants in our survey sample either viewed marketing posts with CWLs or in isolation, and participants in our fMRI sample viewed views posts with either a CWL or a grey filler box. Given that both samples were recruited from populations likely to see cannabis marketing messages as personally relevant, it is possible that the effects of CWLs on sharing intentions, self processing network activity could be explained by distraction. In other words, participants may have viewed all marketing posts as self-relevant when they were able to process them fully but were not able to give them sufficient consideration in our CWL conditions because they were attempting to absorb more information in the limited time afforded for message encoding. This is an important consideration because

real-world processing of marketing posts does not contain any such time constraints and takes place in an information-rich environment. Thus, if the effects of CWLs on self processing and sharing in our study were not driven by the health information presented in CWLs, these labels may not have similar effects outside this controlled setting. To account for this, future studies should test whether the effects of CWLs differ from the effects of other pieces of content likely to exist alongside these posts such as unrelated posts or banner ads for non-cannabis products.

Conclusions

The purpose of this study was to test the effects of pictorial CWLs on decisions to share marketing posts promoting cannabis edibles and on activity in brain regions associated with similar sharing decisions. Results confirmed that the presence of CWLs was associated with reduced intentions to share these posts through social media and other online channels, which aligns with past work suggesting health warnings reduce engagement with social media content. The presence of pictorial CWLs was also associated with less activation of self processing regions. Given past research implicating these regions in sharing decisions, this may suggest that pictorial CWLs reduce sharing intentions partly because they reduce perceptions of self-relevance that inform sharing decisions.

CHAPTER 5: CONCLUSIONS

Dissertation review

Persuasive messages can influence a wide variety of important behaviors (Wakefield, Loken, & Hornik, 2010). Therefore, there is great interest across a variety of fields in gaining a better understanding of how persuasion works. In the past decade, neuroscientists have contributed to our understanding of persuasion by demonstrating that neural processes associated with subjective valuation, self, and social processing are key to the persuasive success (Falk & Scholz, 2018). However, much of this research has focused on neural activity during aggregate level message exposure, making it difficult to determine how different message features influence the underlying processes associated with persuasion. The current dissertation aimed to fill this gap by examining how message features influence the underlying mechanisms associated with persuasion and aimed to situate findings within existing theories of behavior change and persuasion. Specifically, this dissertation aimed to: 1) examine whether neural activity differs based on gain versus loss framed messages and how these differences relate to prospect theory; 2) examine whether neural activity differs based on fear versus humor message appeals and how these differences relate to psychological reactance theory; and 3) examine whether neural activity differs based on messages with warning labels versus no warning labels and how these differences relate to a neural model of information virality. The findings are discussed below, along with a broader discussion about how these findings fit within our current understanding of persuasion and can be used in message design. Finally, this chapter concludes by discussing future directions for this research.

Gain versus loss message frames and prospect theory

First, we examined whether gain- versus loss-framed messages were processed differently in the brain. This research built on previous findings that the gain/loss framing of persuasive messages about sunscreen use affected activity in the vmPFC, which is an established predictor of behavior change (Vezich et al., 2017). Our results aligned with these findings, suggesting the effect of message framing on vmPFC activity can be generalized to multiple message contexts. Gain/loss framing exerted a similar effect on self-reported perceptions of message effectiveness, and we found that vmPFC activity was parametrically modulated by these ratings. These findings align with hypotheses that vmPFC activity during message encoding indicates a process of *self-value integration* (Vezich et al., 2017) or *subjective valuation* (Scholz & Falk, 2018), in which people consider the relevance of message content and weigh the value of incorporating the advocated behavior into their self-concept. Importantly, we also found that message framing only affected vmPFC activation during the encoding of messages that described behaviors with personal outcomes (exercise), but not prosocial outcomes (pro-environmental behavior). This personal/prosocial outcome dimension did not exert direct effects on either vmPFC activity or message effectiveness perceptions, suggesting that the vmPFC's sensitivity to gain/loss framing depends on the nature of behavior being advocated. This is consistent with claims of prospect theory (Kahneman & Taversky, 1973; Tversky & Kahnemann, 1989), which asserts that people are more attentive to information about gains under certain circumstances because they view this information through the lens of personal assessments of risk. Thus, these findings suggest that specific features of message design (gain/loss framing and the nature of message outcomes) can affect activity in the brain in a way that is consistent with a prominent theory of behavioral science.

Fear versus humor appeals and psychological reactance

Second, we examined results from independent survey and fMRI data collection procedures to determine whether differences in the degree to which messages induced self-reported experiences of reactance (Brehm, 1966) could be associated with patterns of activity in the brain. Our findings confirmed that anti-DUIC messages elicited cognitive and affective responses consistent with established models of psychological reactance theory (Dillard & Shen, 2005) in our survey sample, which was itself a novel finding. We also found that anti-DUIC PSAs aligned neural activity across more of the brain than control messages and that messages that aligned activity across more of the brain were also rated as more effective by an independent survey sample. This aligned with past findings that compared responses to structured and unstructured video stimuli (Hasson et al. 2008) and with findings that effective messages associated with behavior change elicit increased ISC than ineffective ones (Imhof et al., 2018). Finding that the structure of both our survey and fMRI datasets aligned with theoretical expectations, we used a rank-order comparison procedure to test whether messages that elicited higher levels of self-reported reactance processes also elicited more ISC in certain *a priori* brain regions of interest. We found that messages that elicited higher levels of anger also elicited more ISC in the anterior insulae (AI), which are associated with negative emotional experiences. We also found that messages that elicited more positive cognitions also elicited activation of the ventral striatum (VS) and ventromedial prefrontal cortex (vmPFC), which have previously been associated with a domain-general value signal in the brain (Scholz et al., 2017). Given that both of these findings compared responses to the same set of messages across two datasets using two different measures, we argue that they suggest certain patterns of neural activity can be used to identify messages that are likely to induce certain reactance-relevant experiences in audiences.

Pictorial warning labels and value-based virality

Third, we analyzed data from both an online survey and an fMRI protocol to test the possibility that a set of proposed pictorial cannabis warning labels (CWLs) might slow the spread of social media posts promoting cannabis products. Our procedure allowed us to test a proposed neural mechanism of sharing decisions, which are supposed to be driven by activity in brain networks associated with self, social, and value processing (Scholz et al., 2020). This work also has relevant implications for public health, as the online sharing of promotional content about cannabis has been described as a critical public health risk (Moreno, 2020) because these posts seldom include information about cannabis health risks and can easily be shared in a way that sidesteps platforms' age restrictions (Moreno et al., 2018; 2022). Our results confirmed that the addition of CWLs to social media posts reduced viewers' intentions to share these posts, both with close friends and with their larger social media networks. Regarding their effects of brain activity, we found that fMRI participants exhibited less activation in self processing regions when viewing posts that were paired with CWLs, but we did not find evidence that the presence of CWLs had any effect on social or value processing regions. Acknowledging some important limitations of the design, we suggested that this might indicate that the effects of CWLs on sharing decisions might be driven in part by effects of CWL content on perceptions of the self-relevance of promotional content.

Discussion

Taken together, these three studies serve as a conceptual bridge connecting the neuroscience of persuasion with the persuasion theories that inform message design. While recent work has shown that theory-informed design features have only small and highly variable effects on message outcomes (O'Keefe & Hoeken, 2021), our results demonstrate that these

features do sometimes affect cognitive and affective processes in ways that are consistent with theory. This is an exciting development because it suggests neuroimaging methods can help inform the conversation about theory-driven message design by observing the message/receiver interactions that give rise to persuasion processes. By observing these interactions in a way that does not depend on participant self-report, researchers can use neuroimaging tools to develop a more mechanistic understanding of persuasion theories. For example, our Chapter 2 findings suggest that the established effects of gain/loss framing on the effectiveness of health messages (Rothman et al., 2020) may be driven by the ways these frames shift people's perceptions of the self-relevance and value of message content. Insights like these provide important texture to our understanding of message features, and they may help explain why manipulation of these features show inconsistent effects on outcomes.

In addition to their promise for theory, our findings suggest neuroimaging tools can play a role in message design. Past research has shown that data obtained from small neural samples can be used to predict the population-level effectiveness of persuasive messages (Falk et al., 2012; 2015b), and our results provide an important extension to this by demonstrating that these tools can also capture the effects of specific message features that generalize across multiple samples. For example, the results of Chapter 4 showed that adding a set of proposed pictorial warning labels to real-world cannabis marketing materials affected both sharing intentions and activity in a key brain network associated with those intentions in two independent samples. Given that social media sharing of these materials is considered a public health concern (Moreno, 2022), these findings demonstrate that fMRI and other neuroimaging tools can be used to pretest not only the persuasive effectiveness of health messages, but also their likely impacts on other relevant behaviors. Similarly, the results of Chapter 3 demonstrated that patterns of

neural activity can be used to identify messages that are more likely to elicit complex affective responses like the ones associated with reactance. As future studies refine the understanding of the neural correlates of these more complex media effects processes, so-called "neural focus groups" (Falk et al., 2010) may be able to anticipate which of these processes are likely to be shared across large segments of a message's audience. This would be particularly useful for screening messages about controversial or stigmatizing topics, which may be difficult to study with traditional focus groups due to social desirability bias (Grimm, 2010).

The results of these studies suggest several other promising opportunities for future work. Most obviously, future studies could test the effects of other common theory-driven design choices, such as narrative vs. nonnarrative or tailored vs. nontailored messages. For these early attempts to connect theoretical claims to neural activity, we suggest that researchers control heterogeneity among messages by developing study-specific stimuli as we have for the study in chapter two. Because real-world messages differ across many dimensions, however, we also urge researchers to test the effects of message features in large samples of messages used in actual persuasive campaigns. Whenever possible, we urge researchers to use campaigns for these studies that have been produced following established best practices (Atkin, 2001) and that have undergone some sort of official outcome evaluation. Comparison of neural activity during message encoding with information from outcome evaluations using a brain-as-predictor style analysis provides an opportunity to build a compelling argument for the connection between message features, cognitive and affective processes, and population-level outcomes.

BIBLIOGRAPHY

Aklin, W. M., Stoeckel, L. E., Green, P. A., Keller, C., King, J. W., Nielsen, L., & Hunter, C. (2020). Commentary: National Institutes of Health (NIH) science of behavior change (SOBC). *Health psychology review*, *14*(1), 193-198.

Alhabash, S., Dong, Y., Moureaud, C., Muraro, I. S., & Hertig, J. B. (2022). Effects of Fear and Humor Appeals in Public Service Announcements (PSAs) on Intentions to Purchase Medications via Social Media. *International Journal of Environmental Research and Public Health*, *19*(19), 12340.

Alia-Klein, N., Gan, G., Gilam, G., Bezek, J., Bruno, A., Denson, T. F., ... & Verona, E. (2020). The feeling of anger: From brain networks to linguistic expressions. *Neuroscience & Biobehavioral Reviews*, *108*, 480-497.

Baek, E. C., Scholz, C., O'Donnell, M. B., & Falk, E. B. (2017). The value of sharing information: a neural account of information transmission. *Psychological science*, *28*(7), 851-861.

Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size affects what people share. *Journal of Marketing Research*, *51*(3), 286-299.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167-218.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412-427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., & Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature neuroscience*, *14*(11), 1468-1474.

Berg, C. J., LoParco, C. R., Cui, Y., Pannell, A., Kong, G., Griffith, L., ... & Cavazos-Rehg, P. A. (2023). A review of social media platform policies that address cannabis promotion, marketing and sales. *Substance Abuse Treatment, Prevention, and Policy*, *18*(1), 35.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Bradley, M. M., Sabatinelli, D., Lang, P. J., Fitzsimmons, J. R., King, W., & Desai, P. (2003). Activation of the visual cortex in motivated attention. *Behavioral neuroscience*, *117*(2), 369.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389.

Brehm, J. W. (1966). *A theory of psychological reactance.* Academic Press.

Brehm, J. W. (1989). Psychological reactance: Theory and applications. *ACR North American Advances*.

Brehm, S. S. (1981). Psychological reactance and the attractiveness of unobtainable objects: Sex differences in children's responses to an elimination of freedom. *Sex Roles*, *7*(9), 937-949.

Briñol, P., & Petty, R. E. (2012). A history of attitudes and persuasion research. *Handbook of the history of social psychology*, 283.

Burns, S. M., Barnes, L. N., Katzman, P. L., Ames, D. L., Falk, E. B., & Lieberman, M. D.(2018). A functional near infrared spectroscopy (fNIRS) replication of the sunscreen persuasion paradigm. *Social Cognitive and Affective Neuroscience*, *13*(6), 628–636.

Cacioppo, J. T., Cacioppo, S., & Petty, R. E. (2018). The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience*, *13*(2), 129–172.

Carliner, H., Brown, Q. L., Sarvet, A. L., & Hasin, D. S. (2017). Cannabis use, attitudes, and legal status in the US: A review. *Preventive medicine, 104,* 13-23.

Cascio, C. N., Scholz, C., & Falk, E. B. (2015). Social influence and the brain: persuasion, susceptibility to influence and retransmission. *Current opinion in behavioral sciences*, *3*, 51-57.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, *129*(3), 564-583.

Chen, G., Taylor, P. A., Shin, Y. W., Reynolds, R. C., & Cox, R. W. (2017). Untangling the relatedness among correlations, Part II: Inter-subject correlation group analysis through linear mixed-effects modeling. *Neuroimage*, *147*, 825-840.

Chiu, V., Hall, W., Chan, G., Hides, L., & Leung, J. (2022). A systematic review of trends in US attitudes toward cannabis legalization. *Substance Use & Misuse*, *57*(7), 1052-1061.

Chua, H. F., Ho, S. S., Jasinska, A. J., Polk, T. A., Welsh, R. C., Liberzon, I., & Strecher, V. J. (2011). Self-related neural response to tailored smoking-cessation messages predicts quitting. *Nature neuroscience*, *14*(4), 426-427.

Cooper, N., Tompson, S., Brook O'Donnell, M., & Emily, B. F. (2015). Brain activity in self-and value-related regions in response to online antismoking messages predicts behavior change. *Journal of Media Psychology*, *27*(3), 93–109.

Cooper, N., Bassett, D. S., & Falk, E. B. (2017). Coherent activity between brain regions that code for value is linked to the malleability of human behavior. *Scientific Reports, 7*(1), 43250.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162–173.

Colonna, R. (2022). Mass media campaigns and media advocacy related to cannabis-impaired driving: a scoping review. *Journal of Substance Use*, 1-9.

Cooper, N., Tompson, S., O'Donnell, M. B., Vettel, J. M., Bassett, D. S., & Falk, E. B. (2018). Associations between coherent neural activity in the brain's value system during antismoking messages and reductions in smoking. *Health Psychology, 37*(4), 375–384.

Covello, V. T., Slovic, P., & Von Winterfeldt, D. (1986). Risk communication: A review of the literature. *Risk Abstracts, 3*, 171-182

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162-173.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature reviews neuroscience*, *10*(1), 59-70.

D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. *Frontiers in human neuroscience*, *7*, 372.

D'Argembeau, A., Jedidi, H., Balteau, E., Bahri, M., Phillips, C., & Salmon, E. (2012). Valuing one's self: medial prefrontal involvement in epistemic and emotive investments in self-views. *Cerebral Cortex, 22*(3), 659-667.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The neuroscientist*, *13*(6), 580-593.

Delhomme, P., De Dobbeleer, W., Forward, S., & Simões, A. (2009). Manual for designing, implementing, and evaluating road safety communication campaigns: Part I. *Brussels: Belgian Road Safety Institute*.

Derzon, J. H., & Lipsey, M. W. (2001). A meta-analysis of the effectiveness of mass-communication for changing substance-use knowledge, attitudes, and behavior. In *Mass media and drug prevention* (pp. 231-258). Psychology Press.

Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication monographs*, *72*(2), 144-168.

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature communications*, *5*(1), 4567.

Do, K. T., & Galván, A. (2015). FDA cigarette warning labels lower craving and elicit frontoinsular activation in adolescent smokers. *Social Cognitive and Affective Neuroscience*, *10*(11), 1484-1496.

Doré, B. P., Tompson, S. H., O'Donnell, M. B., An, L. C., Strecher, V., & Falk, E. B. (2019). Neural mechanisms of emotion regulation moderate the predictive value of affective and value-related brain responses to persuasive messages. *Journal of Neuroscience*, *39*(7), 1293-1300.

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., ... & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PloS one*, *8*(9), e75468.

Eadie, W. F. (2021). *When communication became a discipline*. Rowman & Littlefield.

Elder, R. W., Shults, R. A., Sleet, D. A., Nichols, J. L., Thompson, R. S., Rajab, W., & Task Force on Community Preventive Services. (2004). Effectiveness of mass media campaigns for reducing drinking and driving and alcohol-involved crashes: a systematic review. *American journal of preventive medicine*, *27*(1), 57-65.

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The hippocampus, memory, and place cells: is it spatial memory or a memory space?. *Neuron*, *23*(2), 209-226.

Eisend, M. (2009). A meta-analysis of humor in advertising. *Journal of the Academy of Marketing Science*, *37*, 191-203.

Eisend, M. (2011). How humor in advertising works: A meta-analytic test of alternative models. *Marketing letters*, *22*, 115-132.

Falk, E. B., Berkman, E. T., Mann, T., Harrison, B., & Lieberman, M. D. (2010). Predicting persuasion-induced behavior change from the brain. *Journal of Neuroscience*, *30*(25), 8421-8424.

Falk, E. B., Berkman, E. T., Whalen, D., & Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychology*, *30*(2), 177–185. https://doi.org/10.1037/a0022259

Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From neural responses to population behavior: neural focus group predicts population-level media effects. *Psychological science*, *23*(5), 439-445.

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., & Lieberman, M. D. (2013). Creating buzz: the neural correlates of effective message propagation. *Psychological science, 24*(7), 1234-1242.

Falk, E. B., O'Donnell, M. B., Cascio, C. N., Tinney, F., Kang, Y., Lieberman, M. D., Taylor, S.E., An, L., Resnicow, K., & Strecher, V. J. (2015a). Self-affirmation alters the brain's response to health messages and subsequent behavior change. *Proceedings of the National Academy of Sciences*, *112*(7), 1977–1982. https://doi.org/10.1073/pnas.1500247112

Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual review of psychology*, *69*, 329-356.

Feng, C., Luo, Y. J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human brain mapping*, *36*(2), 591-602.

Finckenauer, J. O. (1982). *Scared straight! and the panacea phenomenon* (pp. 257-257). Englewood Cliffs, NJ: Prentice-Hall.

Fisher, J. T., Hopp, F. R., & Weber, R. (2021). A practical introduction to network neuroscience for communication researchers. *Communication Methods and Measures*, *15*(1), 60-79.

Gardner, L., & Leshner, G. (2016). The role of narrative and other-referencing in attenuating psychological reactance to diabetes self-care messages. *Health communication*, *31*(6), 738-751.

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, *23*(5), 307-318.

Gifford, R. (2011). The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *American Psychologist*, *66*(4), 290–302.

Gilam, G., & Hendler, T. (2017). Deconstructing anger in the human brain. *Social behavior from rodents to humans: Neural foundations and clinical implications*, 257-273.

Grall, C., Tamborini, R., Weber, R., & Schmälzle, R. (2021). Stories collectively engage listeners' brains: Enhanced intersubject correlations during reception of personal narratives. *Journal of Communication*, *71*(2), 332-355.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, *79*(5), 701.

Green, A. E., Mays, D., Falk, E. B., Vallone, D., Gallagher, N., Richardson, A., ... & Niaura, R. S. (2016). Young adult smokers' neural response to graphic cigarette warning labels. *Addictive Behaviors Reports*, *3*, 28-32.

Grewal, J. K., & Loh, L. C. (2020). Health considerations of the legalization of cannabis edibles. *CMAJ*, *192*(1), E1-E2.

Guenther, L., Gaertner, M., & Zeitz, J. (2021). Framing as a concept for health communication: A systematic review. *Health Communication*, *36*(7), 891–899.

Hagger, M. S., Moyers, S., McAnally, K., & McKinley, L. E. (2020). Known knowns and known unknowns on behavior change interventions and mechanisms of action. *Health Psychology Review*, *14*(1), 199-212.

Hagger, M. S., & Weed, M. (2019). DEBATE: Do interventions based on behavioral theory work in the real world?. *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 1-10.

Hall, M. G., Sheeran, P., Noar, S. M., Ribisl, K. M., Boynton, M. H., & Brewer, N. T. (2017). A brief measure of reactance to health warnings. *Journal of Behavioral Medicine*, *40*, 520-529.

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *Neuroimage*, *50*(3), 1313-1319.

Harada, T., Itakura, S., Xu, F., Lee, K., Nakashita, S., Saito, D. N., & Sadato, N. (2009). Neural correlates of the judgment of lying: A functional magnetic resonance imaging study. *Neuroscience research*, *63*(1), 24-34.

Hartman, R. L., & Huestis, M. A. (2013). Cannabis effects on driving skills. *Clinical chemistry*, *59*(3), 478-492.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *science*, *303*(5664), 1634-1640.

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. *Projections*, *2*(1), 1-26.

Helme, D. W., Egan, K. L., Lukacena, K. M., Roberson, L., Zelaya, C. M., McLeary, M. S., & Wolfson, M. (2020). Encouraging disposal of unused opioid analgesics in Appalachia. *Drugs: Education, Prevention and Policy*, 1–9.

Hornik, R., Jacobsohn, L., Orwin, R., Piesse, A., & Kalton, G. (2008). Effects of the national youth anti-drug media campaign on youths. *American Journal of Public Health*, *98*(12), 2229-2236.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion.* Yale University Press.

Huskey, R., Mangus, J. M., Turner, B. O., & Weber, R. (2017). The persuasion network is modulated by drug-use risk and predicts anti-drug message effectiveness. *Social cognitive and affective neuroscience*, *12*(12), 1902-1915.

Imhof, M. A., Schmälzle, R., Renner, B., & Schupp, H. T. (2020). Strong health messages increase audience brain coupling. *NeuroImage*, *216*, 116527.

Imhof, M. A., Schmälzle, R., Renner, B., & Schupp, H. T. (2017). How real-life health messages engage our brains: Shared processing of effective anti-alcohol videos. *Social Cognitive and Affective Neuroscience*, *12*(7), 1188-1196.

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., & Sams, M. (2008). Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *The open neuroimaging journal*, *2*, 14.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 121-140.

Johnson, M. K., Raye, C. L., Mitchell, K. J., Touryan, S. R., Greene, E. J., & Nolen-Hoeksema, S. (2006). Dissociating medial frontal and posterior cingulate activity during self-reflection. *Social cognitive and affective neuroscience*, *1*(1), 56-64.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Kecskemeti, S., Samsonov, A., Velikina, J., Field, A. S., Turski, P., Rowley, H., Lainhart, J. E., & Alexander, A. L. (2018). Robust motion correction strategy for structural MRI in unsedated children demonstrated with three-dimensional radial MPnRAGE. *Radiology*, *289*(2), 509–516.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *science*, *314*(5800), 829-832.

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. *Cerebral cortex*, *18*(9), 1987-1990.

Kreuter, M. W., Sugg-Skinner, C., Holt, C. L., Clark, E. M., Haire-Joshu, D., Fu, Q., Booker, A.C., Steger-May, K., & Bucholtz, D. (2005). Cultural tailoring for mammography and fruit and vegetable intake among low-income African-American women in urban public health centers. *Preventive Medicine, 41*(1), 53–62.

Langrand, J., Dufayet, L., & Vodovar, D. (2019). Marketing of legalised cannabis: a concern about poisoning. *The Lancet, 394*(10200), 735.

Lapinski, M. K., Zhuang, J., Koh, H., & Shi, J. (2017). Descriptive norms and involvement in health and environmental behaviors. *Communication Research, 44*(3), 367–387.

Lee, S. T., & Cheng, I. H. (2010). Assessing the TARES as an ethical model for antismoking ads. *Journal of Health Communication, 15*(1), 55-75.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current opinion in neurobiology*, *22*(6), 1027-1038.

Liang, Y. (Jake), Kee, K. F., & Henderson, L. K. (2018). Towards an integrated model of strategic environmental communication: advancing theories of reactance and planned behavior in a water conservation context. *Journal of Applied Communication*

Lieberman, M. D. (2010). Social cognitive neuroscience. *Handbook of social psychology*, *5*, 143-193.

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self,(situational), and affective processes in medial prefrontal cortex (mPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience & Biobehavioral Reviews*, *99*, 311-328.

Lin, W. J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific reports*, *6*(1), 1-10.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences*, *35*(3), 121-143.

Liu, J., McLaughlin, S., Lazaro, A., & Halpern-Felsher, B. (2020). What does it meme? A qualitative analysis of adolescents' perceptions of tobacco and marijuana messaging. *Public Health Reports*, *135*(5), 578-586.

Liu, J., O'Donnell, M. B., & Falk, E. B. (2021). Deliberation and valence as dissociable components of counterarguing among smokers: evidence from neuroimaging and quantitative linguistic analysis. *Health communication*, *36*(6), 752-763.

Marijuana Policy Project. *State Policies*. Retrieved May 1, 2023, from https://www.mpp.org/states/

Marinello, S. (2023). Social Media Marketing Practices of Illinois Recreational Cannabis Dispensaries in the First Year of Legal Sales: Product Promotions, Branding, and Price Promotions. *Journal of Drug Issues*, 00220426231159542.

Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., & Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, *33*(2), 219-240.

Moreno, M. A. (2022). Concerning Trends in Youth E-Cigarette and Cannabis Use: Dual Use and Social Media Marketing—A Commentary on Roberts et al.(2022). *Journal of Studies on Alcohol and Drugs*, *83*(5), 773-774.

Moreno, M. A., Gower, A. D., Jenkins, M. C., Scheck, J., Sohal, J., Kerr, B., ... & Cox, E. (2018). Social media posts by recreational marijuana companies and administrative code regulations in Washington State. *JAMA Network Open*, *1*(7), e182242-e182242.

Motoki, K., Suzuki, S., Kawashima, R., & Sugiura, M. (2020). A combination of self-reported data and social-related neural measures forecasts viral marketing success on social media. *Journal of Interactive Marketing*, *52*(1), 99-117.

Moyer-Gusc, E., Robinson, M. J., & Mcknight, J. (2018). The role of humor in messaging about the MMR vaccine. *Journal of Health Communication*, *23*(6), 514-522.

Mukherjee, A., & Dubé, L. (2012). Mixing emotions: The use of humor in fear advertising. *Journal of Consumer Behaviour*, *11*(2), 147-161.

Murray, R. J., Schaer, M., & Debbané, M. (2012). Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self-and other-reflection. *Neuroscience & Biobehavioral Reviews*, *36*(3), 1043-1059.

Myers, M. G., Bonar, E. E., & Bohnert, K. M. (2023). Driving under the influence of cannabis, alcohol, and illicit drugs among adults in the United States from 2016 to 2020. *Addictive behaviors*, *140*, 107614.

Nabi, R. L. (2015). Emotional flow in persuasive health messages. *Health communication*, *30*(2), 114-124.

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667-685.

National Academies of Sciences, Engineering, and Medicine. (2017). The health effects of cannabis and cannabinoids: the current state of evidence and recommendations for research.

Nguyen, N., Wong, M., Delucchi, K., & Halpern-Felsher, B. (2022). Adolescents' and young adults' perceptions of risks and benefits differ by type of cannabis products. *Addictive behaviors*, *131*, 107336.

Niederdeppe, J., Kemp, D., Jesch, E., Scolere, L., Greiner Safi, A., Porticella, N., ... & Byrne, S. (2019). Using graphic warning labels to counter effects of social cues and brand imagery in cigarette advertising. *Health Education Research*, *34*(1), 38-49.

Noar, S. M. (2006). A 10-year retrospective of research in health mass media campaigns: where do we go from here?. *Journal of health communication*, *11*(1), 21-42.

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457.

O'Donnell, M. B., Coronel, J., Cascio, C. N., Lieberman, M. D., & Falk, E. B. (2018, May). An fMRI localizer for deliberative counterarguing. In *Social & Affective Neuroscience Society Annual Meeting, Brooklyn, NY*.

O'Keefe, D. J. (2015). Message generalizations that support evidence-based persuasive message design: Specifying the evidentiary requirements. *Health Communication*, *30*(2), 106–113.

O'Keefe, D. J. (2018). Message pretesting using assessments of expected or perceived persuasiveness: Evidence about diagnosticity of relative actual persuasiveness. *Journal of Communication*, *68*(1), 120–142.

O'Keefe, D. J. (2020). Message pretesting using perceived persuasiveness measures: reconsidering the correlational evidence. *Communication Methods and Measures*, *14*(1), 25–37.

O'Keefe, D. J., & Hoeken, H. (2021). Message design choices don't make much difference to persuasiveness and can't be counted on—not even when moderating conditions are specified. *Frontiers in psychology*, *12*, 664160.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Owens, M. M., MacKillop, J., Gray, J. C., Hawkshead, B. E., Murphy, C. M., & Sweet, L. H. (2017). Neural correlates of graphic cigarette warning labels predict smoking cessation relapse. *Psychiatry Research: Neuroimaging*, *262*, 63-70.

Pandey, P., Kang, Y., Cooper, N., O'Donnell, M. B., & Falk, E. B. (2021). Social networks and neural receptivity to persuasive health messages. *Health Psychology*, *40*(4), 285–294.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of research methods in personality psychology*, *1*(2007), 224-239.

Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of life reviews*, *11*(3), 400-435.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of personality and social psychology*, *41*(5), 847.

Phua, J., Lin, J. S. E., & Lim, D. J. (2018). Understanding consumer engagement with celebrity-endorsed E-Cigarette advertising on instagram. *Computers in Human Behavior*, *84*, 93-102.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Quick, B. L., & Bates, B. R. (2010). The use of gain-or loss-frame messages and efficacy appeals to dissuade excessive alcohol consumption among college students: A test of psychological reactance theory. *Journal of health communication*, *15*(6), 603-628.

Quick, B. L., Morgan, S. E., LaVoie, N. R., & Bosch, D. (2014). *Grey's Anatomy* viewing and organ donation attitude formation: Examining mediators bridging this relationship among African Americans, Caucasians, and Latinos. *Communication Research*, *41*(5), 690–716.

Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human communication research*, *39*(1), 47-73.

Rains, S. A., & Turner, M. M. (2007). Psychological reactance and persuasive health communication: A test and extension of the intertwined model. *Human Communication Research*, *33*(2), 241-269.

Ramaekers, J. G. (2018). Driving under the influence of cannabis: an increasing public health concern. *Jama*, *319*(14), 1433-1434.

Randolph, W., & Viswanath, K. (2004). Lessons learned from public health mass media campaigns: marketing health in a crowded media world. *Annu. Rev. Public Health*, *25*, 419-437.

Reboussin, B. A., Wagoner, K. G., Sutfin, E. L., Suerken, C., Ross, J. C., Egan, K. L., ... & Johnson, R. M. (2019). Trends in marijuana edible consumption and perceptions of harm in a cohort of young adults. *Drug and alcohol dependence*, *205*, 107660.

Reynolds-Tylus, T. (2019). Psychological reactance and persuasive health communication: A review of the literature. *Frontiers in Communication*, *4*, 56.

Rice, R. E., & Atkin, C. K. (2009). Public communication campaigns: Theoretical principles and practical applications. In *Media effects* (pp. 452-484). Routledge.

Riddle, M., & Ferrer, R. (2015). The science of behavior change. *APS Observer*, *28*.

Riddle, P. J., Newman-Norlund, R. D., Baer, J., & Thrasher, J. F. (2016). Neural response to pictorial health warning labels can predict smoking behavioral change. *Social Cognitive and Affective Neuroscience*, *11*(11), 1802–1811.

Rimal, R. N., Lapinski, M. K., Turner, M. M., & Smith, K. (2011). The attribute-centered approach for understanding health behaviors: Initial ideas and future research directions. *Studies in Communication Sciences*, *11*(1), 15–34.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software*, *48*, 1-36.

Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: The role of message framing. *Psychological Bulletin*, *121*(1), 3–19.

Rothman, A. J., Desmarais, K. J., & Lenne, R. L. (2020). Moving from research on message framing to principles of message matching: The use of gain-and loss-framed messages to promote healthy behavior. In *Advances in motivation science* (Vol. 7, pp. 43-73). Elsevier.

Russell, C., Rueda, S., Room, R., Tyndall, M., & Fischer, B. (2018). Routes of administration for cannabis use–basic prevalence and related health outcomes: A scoping review and synthesis. *International Journal of Drug Policy*, *52*, 87-96.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755-1758.

Saxe, R., & Kanwisher, N. (2013). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". In *Social neuroscience* (pp. 171-182). Psychology Press.

Schmälzle, R., Häcker, F. E., Honey, C. J., & Hasson, U. (2015). Engaged listeners: shared neural processing of powerful political speeches. *Social cognitive and affective neuroscience*, *10*(8), 1137-1143.

Scholz, C., Jovanova, M., Baek, E. C., & Falk, E. B. (2020a). Media content sharing as a value-based decision. *Current Opinion in Psychology*, *31*, 83-88.

Scholz, C., Baek, E. C., Brook O'Donnell, M., & Falk, E. B. (2020b). Decision-making about broad-and narrowcasting: a neuroscientific perspective. *Media Psychology*, *23*(1), 131-155.

Schmälzle, R., Häcker, F., Renner, B., Honey, C. J., & Schupp, H. T. (2013). Neural correlates of risk perception during real-life risk communication. *Journal of Neuroscience*, *33*(25), 10340-10347.

Schmälzle, R., & Grall, C. (2020). The coupled brains of captivated audiences. *Journal of Media Psychology*.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34.

Shapiro, A. D., & Grafton, S. T. (2020). Subjective value then confidence in human ventromedial prefrontal cortex. *Plos one*, *15*(2), e0225617.

Shen, L. (2015). Antecedents to psychological reactance: The impact of threat, message frame, and choice. *Health communication*, *30*(10), 975-985.

Shen, L., & Coles, V. B. (2015). Fear and psychological reactance. *Zeitschrift für Psychologie*.

Skalski, P., Tamborini, R., Glazer, E., & Smith, S. (2009). Effects of humor on presence and recall of persuasive messages. *Communication Quarterly*, *57*(2), 136-153.

Shen, L. (2011). The effectiveness of empathy-versus fear-arousing antismoking PSAs. *Health communication*, *26*(5), 404-415.

Siegel, J. T., Alvaro, E. M., Crano, W. D., Gonzalez, A. V., Tang, J. C., & Jones, S. P. (2010). Passive-positive organ donor registration behavior: A mixed method assessment of the IIFF Model. *Psychology, Health & Medicine*, *15*(2), 198–209.

Slavin, M. N., & Earleywine, M. (2019). Effects of messaging and psychological reactance on marijuana craving. *Substance use & misuse*, *54*(14), 2359-2367.

Smith, M. J. (1977). The effects of threats to attitudinal freedom as a function of message quality and initial receiver attitude. *Communications Monographs*, *44*(3), 196-206.

Snyder, L. B. (2007). Health communication campaigns and their impact on behavior. *Journal of nutrition education and behavior*, *39*(2), S32-S40.

So, J. (2013). A further extension of the extended parallel process model (E-EPPM): Implications of cognitive appraisal theory of emotion and dispositional coping style. *Health communication*, *28*(1), 72-83.

Sorella, S., Grecucci, A., Piretti, L., & Job, R. (2021). Do anger perception and the experience of anger share common neural mechanisms? Coordinate-based meta-analytic evidence of similar and different mechanisms from functional neuroimaging studies. *NeuroImage*, *230*, 117777.

Speitel, C., Traut-Mattausch, E., & Jonas, E. (2019). Functions of the right dlPFC and right TPJ in proposers and responders in the ultimatum game. *Social Cognitive and Affective Neuroscience*, *14*(3), 263-270.

Spindle, T. R., Bonn-Miller, M. O., & Vandrey, R. (2019). Changing landscape of cannabis: novel products, formulations, and methods of administration. *Current opinion in psychology*, *30*, 98-102.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance. *Zeitschrift für Psychologie*.

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, *66*(4), 742–752.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin Books.

Thrasher, J. F., Brewer, N. T., Niederdeppe, J., Peters, E., Strasser, A. A., Grana, R., & Kaufman, A. R. (2019). Advancing tobacco product warning labels research methods and theory: a summary of a grantee meeting held by the US National Cancer Institute. *Nicotine and Tobacco Research*, *21*(7), 855-862.

Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive brain research*, *23*(1), 34-50.

Tversky, A., & Kahneman, D. (1989). Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers* (pp. 81-126). Berlin, Heidelberg: Springer Berlin Heidelberg.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*(3), 564-584.

Varava, K. A., & Quick, B. L. (2015). Adolescents and movie ratings: Is psychological reactance a theoretical explanation for the forbidden fruit effect?. *Journal of Broadcasting & Electronic Media*, *59*(1), 149-168.

Ventresca, M., & Elliott, C. (2022). Cannabis edibles packaging: Communicative objects in a growing market. *International Journal of Drug Policy*, *103*, 103645.

Vezich, I. S., Falk, E. B., & Lieberman, M. D. (2016). Persuasion neuroscience: New potential to test dual-process theories. In E. Harmon-Jones & M. Inzlicht (Eds.), *Social Neuroscience: Biological approaches to social psychology* (pp. 34–58). Routledge.

Vezich, I. S., Katzman, P. L., Ames, D. L., Falk, E. B., & Lieberman, M. D. (2017). Modulating the neural bases of persuasion: why/how, gain/loss, and users/non-users. *Social cognitive and affective neuroscience*, *12*(2), 283-297.\

Wakefield, M. A., Loken, B., & Hornik, R. C. (2010). Use of mass media campaigns to change health behaviour. *The lancet*, *376*(9748), 1261-1271.

Wang, A. L., Lowen, S. B., Romer, D., Giorno, M., & Langleben, D. D. (2015). Emotional reaction facilitates the brain and behavioural impact of graphic cigarette warning labels in smokers. *Tobacco control*, *24*(3), 225-232.

Ward, N. J., Finley, K., Townsend, A., & Scott, B. G. (2021). The effects of message threat on psychological reactance to traffic safety messaging. *Transportation research part F: traffic psychology and behaviour*, *80*, 250-259.

Weber, R., Huskey, R., Mangus, J. M., Westcott-Baker, A., & Turner, B. O. (2015). Neural predictors of message effectiveness during counterarguing in antidrug campaigns. *Communication Monographs*, *82*(1), 4-30.

Weber, R., Fisher, J. T., Hopp, F. R., & Lonergan, C. (2018). Taking messages into the magnet: Method–theory synergy in communication neuroscience. *Communication Monographs*, *85*(1), 81-102.

Weinberger, M. G., & Gulas, C. S. (1992). The impact of humor in advertising: A review. *Journal of advertising*, *21*(4), 35-59.

Wicklund, R. A. (1974). *Freedom and reactance.* Lawrence Erlbaum.

Williamson, L. D., Reynolds-Tylus, T., Quick, B. L., & Shuck, M. (2017). African-Americans' perceptions of organ donation: 'simply boils down to mistrust!' *Journal of Applied Communication Research, 45*(2), 199–217.

Willoughby, J. F., Hust, S. J., Li, J., & Couto, L. (2023). Exposure to pro and anti-cannabis social media messages and teens' and college students' intentions to use cannabis. *Health communication*, 1-12.

Windle, S. B., Eisenberg, M. J., Reynier, P., Cabaussel, J., Thombs, B. D., Grad, R., ... & Filion, K. B. (2021). Association between legalization of recreational cannabis and fatal motor vehicle collisions in the United States: an ecologic study. *Canadian Medical Association Open Access Journal*, *9*(1), E233-E241.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs*, *59*(4), 329-349.

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health education & behavior*, *27*(5), 591-615.

Worchel, S., & Brehm, J. W. (1970). Effect of threats to attitudinal freedom as a function of agreement with the communicator. *Journal of Personality and Social Psychology*, *14*(1), 18.

World Health Organization. (2016). The health and social effects of nonmedical cannabis use.

Wu, J., Origgi, J. M., Ranker, L. R., Bhatnagar, A., Robertson, R. M., Xuan, Z., ... & Fetterman, J. L. (2023). Compliance With the US Food and Drug Administration's Guidelines for Health Warning Labels and Engagement in Little Cigar and Cigarillo Content: Computer Vision Analysis of Instagram Posts. *JMIR infodemiology*, *3*, e41969.

Wundersitz, L., Hutchinson, T., & Woolley, J. (2010). Best practice in road safety mass media campaigns: A literature review. *Social psychology*, *5*, 119-186.

Yang, D. Y. J., Rosenblau, G., Keifer, C., & Pelphrey, K. A. (2015). An integrative neural model of social perception, action observation, and theory of mind. *Neuroscience & Biobehavioral Reviews*, *51*, 263-275.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al.(2009). *Perspectives on psychological science*, *4*(3), 294-298.

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, *48*(9), 2658-2664.

Zhao, X., Roditis, M. L., & Alexander, T. N. (2019). Fear and humor appeals in "The Real Cost" campaign: Evidence of potential effectiveness in message pretesting. *American journal of preventive medicine*, *56*(2), S31-S39.