UNIVERSITY OF CALIFORNIA

Santa Barbara

Moral Idiosynchrony: Variability in Naturalistic Complexity
Modulates Intersubject Representational Similarity in Moral Cognition

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Communication

by

Frederic René Hopp

Committee in charge:

Professor René Weber, Chair

Professor Scott Grafton

Professor Scott Reid

December 2021

The dissertation of Frederic René Hopp is approved.

Scott Grafton

Scott Reid

René Weber, Committee Chair

December 2021

Moral Idiosynchrony: Variability in Naturalistic Complexity

Modulates Intersubject Representational Similarity in Moral Cognition

by

Frederic René Hopp

ACKNOWLEDGEMENTS

"Standing on the shoulders of giants" is a common metaphor to illustrate the temple upon which science is built. This dissertation marks one brick in this temple. A brick that was carried through a global pandemic, across three countries, and two continents. This brick would not exist had I not received the invaluable, consistent, and deep support from my academic and non-academic giants. There simply is not enough room to thank each and every one of you, so I will briefly highlight those who had the most profound impact on the herein developed intellectual work.

During my undergraduate studies at the University of Mannheim, working as a research assistant with Professor Peter Vorderer first sparked my interest in thinking deeper about how and why humans spend time with entertainment products and fictional narratives. Professor Vorderer introduced me to academic conferences, scientific working, and is sort of responsible for dragging me away from my hobbies, the corporate world, and into the long and dimly lit halls of science. During this time, I was extremely glad to spend a dark German winter abroad in sunny Southern California, where I had my first encounter with the University of California, Santa Barbara (UCSB), its wonderful Department of Communication, and the Media Neuroscience Lab (MNL), to which I subsequently returned as a graduate student. Returning to the MNL for my MA/PhD training and leaving my home, my friends, and family was motivated by countless wonderful individuals at UCSB.

First and foremost, I have to thank Professor René Weber, the principal investigator of the MNL, the supervisor of my Master thesis and dissertation, and the force-of-nature that convinced me to pursue a doctoral degree with all my heart and energy. Working with René has not only been tremendously inspiring, instructive, and rewarding, but his persistent drive

to push me onto scientific problems that seemed intractable shaped my perseverance, curiosity, and belief that every question has an answer. René has been a wonderful mentor and friend, who always had time and energy for “5-minute” discussions, many of which produced ideas upon which this work is built. Moreover, I would like to thank my doctoral committee members: Professor Scott Reid, for all the engaging discussions on evolutionary approaches to media psychology, and Professor Scott Grafton, for taking a communication graduate student into the wild west of functional neuroimaging.

Second, there is a fellowship of individuals who accompanied and helped me thrive at UCSB. Jacob Fisher, who I consider my Samwise Gamgee, resembled everything one could hope for in a co-worker and friend. Whether it was the shared frustration over algebraic equations on “small graphs” and late night paper crunches, the countless discussions on whiteboards, or the always anticipated relief of Pizza Saturdays at the Fisher’s, Jacob could always be counted on. Next, there were the many early morning surf sessions, late night conference drinks, and data collection sessions with Richard Huskey, who played a major role in convincing me to join graduate school at UCSB. In addition, my fellow lab members Chelsea Lonergan, Musa Malik, and Yibei Chen were not only brilliant collaborators who always had an open ear to listen to my research ideas, but were there for me when things got rough. Furthermore, I would like to express my sincere gratitude towards Andreas Boschke, Jeff Oakes, and J. Michael Mangus, who invested great time and patience in teaching me invaluable skills about system architecture and high performance computing that made this work possible. Moreover, I extend my thanks to Sara Miller McCune and the McCune foundation for their generous funding of my doctoral work. Likewise, I have to thank Tricia Taylor for her wonderful graduate advising, reminding me of deadlines, and saving me countless times when I missed them. I would also like to thank

Professor Ronald Rice and Professor Joseph Walther, for numerous letters of support, good discussions, and their scholarly friendship.

In my personal life, I wish to thank Isaac Mackey and Cole Hawkins, my one-of-a-kind roommates, for all bike rides, ocean swims, sailing trips, math and coding tutoring, and so much more. Furthermore, I wish to thank my beloved parents, Reinhard and Barbara. Words simply cannot capture my gratitude for all the love and support you have given me; for raising me to pursue my passion; and for the many wonderful moments and memories that carried me through difficult times. Finally, I wish to thank my soulmate and my most loving supporter, Christiane. Thank you for all your care and time, for your patience, for getting me through the pandemic, for moving across Europe with me, and for your infinite love.

VITA OF FREDERIC RENÉ HOPP
DECEMBER 2021

EDUCATION

Bachelor of Arts in Media & Communication Studies, University of Mannheim, January 2016

Master of Arts in Communication, University of California Santa Barbara, December 2018

Doctor of Philosophy in Communication, University of California Santa Barbara, December 2021 (exp.)

PROFESSIONAL EMPLOYMENT

August 2021: Assistant Professor, Amsterdam School of Communication Research, University of Amsterdam

September 2018 - July 2020: Graduate Student Researcher, University of California Santa Barbara

July 2020 - August 2020: Associate Instructor, University of California Santa Barbara

September 2016 - August 2017: Graduate Teaching Assistant, University of California Santa Barbara

January 2013 - August 2016: Research Assistant, University of Mannheim

PUBLICATIONS

Malik, M., Hopp, F. R., Chen, Y., & Weber, R. (2021). Does regional variation in pathogen prevalence predict the moralization of COVID-19 in online news? *Journal of Language and Social Psychology.*

Hopp, F. R., & Weber, R. (2021). Rejoinder: How methodological decisions impact the validity of moral content analyses. *Communication Monographs.*

Hopp, F. R., & Weber, R. (2021). Reflections on extracting moral foundations from media content. *Communication Monographs.*

Fisher, J.T., Lonergan, C., Hopp, F.R., & Weber, R. (2021) Media entertainment, flow experiences, and the synchronization of audiences. In P., Vorderer, & C., Klimmt (Eds.), *Oxford Handbook of Entertainment Theory*. Oxford, UK: Oxford University Press.

Rohm, S., Hopp, F. R., & Smit, E.G. (2021). Exposure to serial audiovisual narratives increases empathy via vicarious interactions. *Media Psychology.*

Fisher, J. T., Hopp, F. R., & Weber, R. (2020). A practical introduction to network neuroscience for communication researchers. *Communication Methods and Measures*, 1–20.

Hopp, F.R., Fisher, J., & Weber, R. (2020). A graph-learning approach for detecting moral conflict in movie scripts. *Media and Communication*, *8*(3), 164–179.

Hopp, F.R., Fisher, J., Cornell, D., Huskey, R., & Weber, R. (2020). The extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*.

Hopp, F.R., Fisher, J., & Weber, R. (2020). Dynamic transactions between news frames and sociopolitical events: An integrative, hidden markov model approach. *Journal of Communication*, *70*(3), 335–355.

Hopp, F.R., & Weber, R. (2020) The state-of-the-art and future of fMRI methodology in communication research. In R., Weber & K., Floyd (Eds.), *Handbook of Communication Science and Biology*. New York, NY: Routledge.

Weber, R., & Hopp, F.R. (2020). Moral emotions and conflict motivate actions. *Insights– Consumer Neuroscience in Business*, *30*, 12–13.

Weber, R., Hopp, F.R., & Fisher, J.(2020). The moral narrative analyzer (MoNA): A platform for extracting moral emotions and conflict from messages at scale. *Neuromarketing Yearbook 2020*. Neuromarketing Science & Business Association (NMSBA).

Fisher, J., Hopp, F.R., & Weber, R. (2019). Modality-specific effects of perceptual load in multimedia processing. *Media and Communication*, *7*(4), 149–165.

Hopp, F.R., Schaffer, J., Fisher, T., Cornell, D., & Weber, R. (2019). iCoRe: The GDELT interface for the advancement of communication research. *Computational Communication Research*, *1*(1), 13–44.

Weber, R., Mangus, J., M., Huskey. R., Hopp, F.R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *2*(2-3), 119–139.

Weber, R., Fisher, J.T., Hopp, F.R., Lonergan, C. (2017). Taking messages into the magnet: Method-theory synergy in media neuroscience research. *Communication Monographs*, *84*, 1–22.

Weinmann, C., Roth, F. S., Schneider, F. S., Krämer, T., Hopp, F.R., & Vorderer, P. (2017). I don’t care about politics, I just like that guy! Affective disposition and political attributes in information processing of political talk shows. *International Journal of Communication*, *11*, 3118–3140.

Schneider, F. M., Zwillich, B., Bindl, M., Hopp, F.R., Vorderer, P.,Reich, S. (2017). Social media ostracism: The effects of being excluded online. *Computers in Human Behavior, 73,* 385–393.

Roth, F. S., Weinmann, C., Schneider, F. S., Hopp, F.R.,Bindl, M. J., & Vorderer, P. (2017).Curving entertainment: The curvilinear relationship between hedonic and eudaimonic experiences while watching a political talk show and its implications for information processing. *Psychology of Popular Media Culture*, *6*, 1–19.

Roth, F., Weinmann, C., Schneider, F., Hopp, F.R., & Vorderer, P. (2014). Seriously entertained: Antecedents and consequences of hedonic and eudaimonic entertainment experiences with political talk shows on TV. *Mass Communication and Society, 17*(3), 379–399.

AWARDS

2021: James J. Bradac Award for Outstanding Graduate Research, University of California Santa Barbara

2020: George D. McCune Dissertation Fellowship, University of California Santa Barbara

2020: Top Paper Award, National Communication Association – Mass Communication Division

2020: Top Paper Award, International Communication Association – Computational Methods Division

2020: Top Paper Award, International Communication Association – Communication Science & Biology

2019: Top Paper Award, National Communication Association – Communication and Social Cognition

2019: Top Paper Award, International Communication Association – Computational Methods Division

2018: Article of the Year Award – Communication Methods and Measures

2018: Top Paper Award, National Communication Association – Mass Communication Division

2018: Top Paper Award, International Communication Association – Communication Science & Biology

2015: PROMOS Scholarship – German Academic Exchange Service

ABSTRACT

Moral Idiosynchrony: Variability in Naturalistic Complexity Modulates Intersubject Representational Similarity in Moral Cognition

by

Frederic René Hopp

In daily life, moral judgments are embedded in dynamic, complex, and contextualized environments. As we reason about morally right or wrong behaviors, our personal history shapes how we judge who did what to whom, where, when, and why. Yet, surprisingly little is known about how individual differences in moral dispositions modulate shared neural response patterns when processing increasingly complex moral scenarios. Consequently, we herein examine brain-behavior-trait coherence in moral cognition across three datasets of increasing naturalistic complexity. Applying intersubject representational similarity analysis, we demonstrate how between-subject variability in moral dispositions modulates similarity in neural responses when processing decontextualized moral vignettes, auditory movie summaries, political attack advertisement, soap opera clips, and full-length movies. Our approach highlights how brain-behavior-trait relationships during moral cognition are shaped by paradigm choice, and provides a reference for conducting research at the intersection of socio-moral cognition, communication science, and naturalistic neuroimaging.

Table of Contents

Introduction

In daily life, the human mind integrates myriad sources of sensory information when judging an act as morally right or wrong. Did the cook knowingly serve rotten food or was it an accident? Is torture justifiable when it saves innocent lives? Do I yell at the referee when his false call leads to the win of my son’s hockey team? Pondering on these questions reveals that moral judgment is influenced by a host of exogenous information embedded in our environment, including our social context (Yudkin et al., 2021) and cues about a perpetrator’s mental state (Young & Saxe, 2008), character (Inbar et al., 2012; Uhlmann et al., 2015), group affiliation (Eriksson et al., 2019; Voelkel & Brandt, 2019), race and gender (Hester & Gray, 2020), and intention (Koster-Hale et al., 2013; Schaich Borg et al., 2006; Young & Saxe, 2009). In addition, moral judgment is also shaped by endogenous information that reflects our current homeostatic states (Crockett et al., 2008), past socialization and environmental pressures (Garten et al., 2019; Graham et al., 2009; van Leeuwen et al., 2014), and future aspirations (Berman & Kupor, 2020). The integration of both exogenous and endogenous information allows us to meaningfully judge if an observed behavior is morally right or wrong (Schein, 2020).

In view of this multidimensional nature, recent calls (Hester & Gray, 2020; Redcay & Moraczewski, 2020; Schein, 2020) debate the ecological validity of highly controlled, decontextualized experimental paradigms that form the pillars of moral psychology (Greene et al., 2001; Greene & Haidt, 2002; Shenhav & Greene, 2010; Wasserman et al., 2017; Young & Dungan, 2012). While these paradigms have instrumentally driven the “localizationist” objective of assigning specific aspects of moral judgment to discrete brain regions, the abstract and impoverished nature of trolley problems and vignettes of “raceless, genderless strangers” (Hester & Gray, 2020) rarely simulate the dynamic, complex, and multimodal information that shapes everyday moral cognition. To better capture real-life sensory experience, cognate areas in the neurosciences have started using more “naturalistic” paradigms, which generally refer to “rich, multimodal dynamic stimuli that represent our daily lived experience, such as film clips, TV advertisements, news items, and spoken narratives, or that embody relatively unconstrained interactions with other agents, gaming environments, or virtual realities.” (Sonkusare et al., 2019). Particularly when contextual information matters (McIntosh, 2004), naturalistic stimuli can greatly complement controlled, abstract task-based paradigms. Evidence for the contextual embeddedness of moral information is ample (Schein, 2020), yet naturalistic paradigms within moral psychology remain scarce (Adebimpe et al., 2019; Pegado et al., 2018). This slow adoption may be attributed to the challenge of (*i*) defining and capturing the latent, complex dynamics of morally relevant content embedded in naturalistic stimuli (Hopp, Fisher, & Weber, 2020; Hopp, Fisher, Cornell, et al., 2020; Weber et al., 2018) and (*ii*) jointly modelling the exogenous and endogenous information brought forth by combining naturalistic stimuli with neural activity and behavior (Turner et al., 2019).

We introduce a research framework for tackling these challenges. We start by highlighting the promise of character-driven narratives as naturalistic stimuli for studying how exogenous and endogenous information shape neural representations of moral cognition. Thereafter, we present content-analytic methodologies for extracting morally relevant information from narratives and discuss techniques for continuously interrogating moral cognition during narrative processing. Subsequently, we apply intersubject representational similarity analysis (IS-RSA) to examine shared structure in neural representations of moral information across three datasets of increasing naturalistic complexity. Specifically, we demonstrate how between-subject variability in moral traits and behavior drives similarity in neural response patterns when processing experimentally-controlled moral vignettes and narrated movie summaries; naturalistic audiovisual political attack advertisements; clips from a popular U.S. TV soap opera, and full-length movies. We envision that our proposed framework pushes scientific progress beyond moral judgment, providing a reference for conducting advanced standardized research at the intersection of socio-moral cognition, communication science, and naturalistic neuroimaging.

Narratives as Paradigm in Moral Neuroscience

Narratives have become essential stimuli for advancing neuroscientific research (Nastase, Liu, Hillman, Zadbood, et al., 2020; Willems et al., 2020). The inherent structure of narratives enables examinations of how the brain processes complex streams of sensory information, including social networks among characters and dynamic events that cause and are caused by multiple agents' actions (Baldassano et al., 2018). From a computational perspective, narratives simulate humans' natural experience by presenting arrays of events in a temporally sequenced, causally related order (Mar & Oatley, 2008), thereby providing a more ecologically valid assessment of how the human brain operates "in the wild" (Speer et al., 2009). At the same time, narratives can be constructed and edited to isolate or control some aspect of scientific interest (Chang et al., 2020; Tamborini et al., 2013; Zillmann & Cantor, 1977), making them amenable for use in experimental settings. In what follows, we argue that narratives are especially powerful tools for advancing the study of moral cognition (Kelly & O'Connell, 2020). While narratives exists in different types and forms, we understand narratives as "cohesive and coherent stor[ies] with an identifiable beginning, middle, and end that provide information about scene, characters, and conflict; raise unanswered questions or unresolved conflict; and provide resolution" (Hinyard & Kreuter, 2007, p. 778). Paradigmatic examples are novels, screenplays, and movies, whereas we do not consider computer code or instruction manuals as narratives.

Narratives Feature Contextualized Moral Information

In many ways, narratives provide a continuous array of contextually embedded moral cues. Within this input stream, characters are the central carrier of morally relevant information (Tamborini, 2013; Zillmann, 2002). Characters are frequently driven by moral and immoral motivations to achieve their goals (Eden et al., 2021), and are placed in settings that foster or constraint moral behavior. In addition, characters experience consequences for their moral actions in the form of rewards and punishments (Weber, Tamborini, et al., 2008), and hence serve as important models for vicarious moral learning (Bandura, 2001). Moreover, characters frequently face realistic moral conflicts and dilemmas (Hopp, Fisher, & Weber, 2020), where either some moral norm must be violated to uphold other norms, or egoistic and altruistic motivations clash (Tamborini & Weber, 2020). In contrast to abstract agents, characters inherit many human features that modulate moral judgment in real-life, spanning visual (e.g., age, sex, race), social (e.g., power, status), mental (e.g., empathy, psychopathy), and relational (e.g., ally versus enemy) attributes (for an overview, see Hester & Gray, 2020). Screenwriters and directors often make the moral makeup of characters hypersalient to facilitate engaging storytelling (McKee, 1997). In this sense, narratives provide key contextual information that exists in the background of real-word moral decision-making. Thus, narratives may serve as an even richer sampling space of moral scenarios in contrast to everyday experience. At the same time, not all narratives prominently feature moral information, while few might not contain any events of moral relevance. However, mounting evidence suggests that universally appealing narratives that engage large numbers of culturally-diverse audiences all weave a tapestry of moral truisms that are recognized across the globe (Weber et al., 2013). In this sense, morally relevant narratives might not only be useful tools for understanding human moral cognition, but indeed may unlock some of the mysteries undergirding compelling storytelling.

Narrative Processing Elicits Moral Judgment

Media psychologists have long recognized that the judgment of characters' moral actions is a central, continuous process undergirding narrative exposure (Weber et al., 2008), character perception (Eden et al., 2011, 2015, 2017; Kleemans et al., 2017; Krakowiak & Tsay-Vogel, 2015), and story appeal (Raney, 2002, 2004; Weber et al., 2008). According to Affective Disposition Theory (ADT; Zillmann & Cantor, 1976), narrative consumers act as "untiring moral monitors" (Zillmann, 2000, p. 54), who continually evaluate the moral actions of characters. As the experienced story unfolds, audiences form affective (i.e., positively or negatively valenced) dispositions towards characters based on the perceived moral rightness or wrongness of their actions: Positive affective disposition are developed towards characters whose behavior conforms to audiences' moral sensitivities, whereas negative affective dispositions are formed towards characters whose actions violate moral norms. In turn, audiences frequently select and enjoy stories where positively evaluated characters are rewarded and negatively appraised characters are punished (Raney, 2002, 2004; Weber et al., 2008). Moreover, stereotypical, visual cues associated with heroes and villains evoke character-consistent moral judgments *absent* character behaviors (Grizzard et al., 2018), resembling person-centered moral judgments that emphasize someone's *identity* over their actions (Uhlmann et al., 2015). Furthermore, these character-schema activations are magnified by the presence of an opposing character, and schema-based moral judgments can bias approbation of behaviors and subsequent moral judgments of characters (Grizzard et al., 2018). In addition, there exist networked interdependencies that guide moral judgment during narrative exposure, such that affective dispositions formed toward one character are interdependent with affective dispositions formed toward others (Grizzard et al., 2020). Although a full review of how moral judgment operates during narrative exposure is beyond the scope of this work (for excellent summaries, see Tamborini, 2013; Zillmann & Vorderer, 2000), our point is that narrative exposure *does* elicit moral judgments that mirror those in daily life.

Narratives Emphasize Individual Variation in Moral Intuition

People vary in how they reach moral judgments in many ways. At the extreme, "even when identical actors perform identical actions in identical situations, moral evaluations can differ widely depending on *who* is doing the judging" (Schein, 2020; p. 210; emphasis added). Over the last decades, moral psychology has produced numerous insights how individual differences or situational factors modulate moral judgment (Hester & Grey, 2020). Narratives hold great potential to connect to this research, highlighting how the processing of dynamic, exogenous moral information is modulated by endogenous moral sensitivities. For example, individuals' moral intuition salience–the relative weight attributed to moral norms as a result of socialization processes (Graham et al., 2011) – predict narrative genre preferences (Bowman et al., 2012), character perception (Grizzard et al., 2018; Grizzard et al, 2019), story appeal (Tamborini et al., 2013), and even intercoder reliabilities in moral content codings (Weber et al., 2018). In addition, the dynamic and contextual nature of narratives may reveal individual differences in moral cognition that remain undetected when using static, decontextualized stimuli: Someone might strongly endorse principles of loyalty at an abstract level, while fully approving a character who commits an act of whistleblowing for scientific fraud (Schein, 2020). Here, the narrative context reveals that loyalty to *whom* is key. From a cognitive perspective, evidence suggests that the brain may simply be more “tuned” to naturalistic than artificial stimuli: for example, the addition of natural, biological motion to facial stimuli increases the strength of cortical responses even in those regions, such as the fusiform gyrus, that classically respond to static faces (Schultz & Pilz, 2009). Accordingly, individual differences at both the neural and behavioral level may be accentuated when increasing the naturalistic complexity of moral stimuli. Indeed, mounting research demonstrates that naturalistic, in contrast to controlled, static paradigms, outperform the detection of individual differences (Chang et al., 2021; Chen et al., 2019; Finn et al., 2018, 2020; Finn & Bandettini, 2021; Vanderwal et al., 2017).

Extracting Moral Information From Narrative Content and Cognition

Narratives contain a rich moral sampling space for interrogating human moral cognition. Our understanding of how the human brain processes, integrates, and responds to morally relevant narratives is therefore fundamentally constrained by our understanding of the contextually embedded moral content in narratives (Hopp, Fisher, & Weber, 2020; Hopp, Fisher, Cornell, et al., 2020; Weber, 2008; Weber et al., 2018) and our confidence that observed neural responses during narrative perception reflect underlying moral computations (Krakauer et al., 2017). In the following, we argue that overcoming both challenges requires (*i*) a methodical, theoretically-informed content analysis of moral content, and (*ii*) behavioral measurements that tap into moral cognition during narrative processing.

The Seven Pillars of Moral Content

Extracting latent, contextually embedded moral content is a challenging task, calling for methodical, fine-grained, and theoretically informed content analysis (Weber et al., 2018). Failure to adequately embrace the complexity of moral content would otherwise re-introduce several limitations that dynamic, multimodal naturalistic stimuli aim to overcome (Hopp & Weber, 2020). A holistic account of moral content in narratives begs the question: *Who*, does *what*, *when*, to *whom*, with what *reason* and what *intention*, and with what *consequence*? Capturing the components of this "moral arc" in narratives is instrumental for making informed predictions about neural responses to morally salient events. Moreover, while the detection of narrative agents (*who* and *whom*) at a particular time point (*when*) is a rather manifest task, defining and capturing what constitutes a *moral* act (*what*), and determining the reasons and intentions (*why*) and outcomes (*consequence*) of this act requires more nuanced, theoretically-informed definitions.

The majority of research in the communication literature examining the types of moral acts that are represented in narratives has relied on the Model of Intuitive Morality and Exemplars (MIME, Eden et al., 2021; Tamborini, 2013; Tamborini & Weber, 2020). The MIME defines morality by largely relying on Moral Foundations Theory (MFT; Haidt, 2007; Graham et al., 2012), which conceptualizes morality with regard to five moral foundations: The *individualizing* foundations spanning *Care/harm* (involving intuitions of sympathy, compassion, and nurturance) and *Fairness/cheating* (including notions of rights and justice), as well as the *binding foundations* spanning *Loyalty/betrayal* (supporting moral obligations of patriotism and "us versus them" thinking), *Authority/ subversion* (including concerns about traditions and maintaining social order), and *Sanctity/degradation* (including moral disgust and spiritual concerns related to the body). A sixth foundation—*Liberty/oppression* (an intuition about the feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty)—is still under consideration (see http://moralfoundations.org). Guided by the MIME, Weber and colleagues (2018) demonstrated that a large crowd of human coders, appropriately trained in MFT and using an intuitive annotation tool, yield the most reliable and valid codings of moral foundations in textual narratives. While Weber and colleagues did not specifically code for the motives behind moral actions, several validated measures exist to relate intentions of agents and targets (e.g., benefit or harm to self (agent) or others (target); NTVS, 1996). The coding of characters' consequences has mostly relied on ratings of outcome valence (e.g., from extremely good to extremely bad; Tamborini et al., 2013; Weber et al., 2008). In addition to manual content annotations, recent research has utilized advancements in natural language processing and social network analysis to computationally detect moral communities in character networks of movie scripts (Hopp, Fisher, & Weber, 2020), and automated the detection of moral foundations in textual corpora (Hopp, Fisher, Cornell, et al., 2020).

Triangulating Moral Cognition During Narrative Processing

Defining the morally salient events embedded in narrative is a necessary, but insufficient step for understanding moral cognition during narrative processing. Even if content features labelled as "morally relevant" elicit neural responses in brain networks associated with moral cognition, the myriad dynamic and sensory inputs received during narrative perception presage reverse inference errors (Poldrack, 2011). In addition, neural responses alone cannot reveal the semantic and valenced appraisal of morally relevant characters and events, which is crucial for understanding the underlying moral computation that preceded neural activation patterns. Hence, neural responses must be complemented with behavioral measurements that tap into these moral appraisals (Krakauer et al., 2017; Turner et al., 2019). These behavioral measurements are typically static (following narrative exposure) or dynamic (during narrative exposure). Static measures may retrospectively interrogate perceived character morality (Grizzard et al., 2019) and character outcomes (Weber et al., 2008). In contrast, dynamic measurements provide time-locked, moment-by-moment assessments of select stimuli features. For example, continuous response measurements (CRM; Biocca et al., 1993) have been used to dynamically rate stimuli's valence and arousal (Nummenmaa et al., 2012), but may equally be used to continually index moral judgments of characters. Furthermore, think-aloud and free-recall paradigms have been administered to survey individuals' thoughts and interpretations during narrative exposure (Nguyen et al., 2019; Weber, 2008). Responses can subsequently be content-analyzed to determine moral appraisals. In addition, individuals' facial expressions may reveal several dimensions of their emotional states (Ekman, 1993). As recent advancements in the field of affective computing have yielded impressive progress in automatically detecting facial expressions from videos (Cheong et al., 2021), analyzing facial expressions during narrative consumption may provide novel insights into elicited affective experiences (Chang et al., 2021, Mangus, 2016). Although the application of static or dynamic measurements depends on the research question, a particular strength of dynamic measures is their shared temporal resolution with neural recordings (i.e., continuous, moment-by-moment), which affords a more straightforward mapping between behavioral and neural responses.

Joint Modeling of Moral Information Across Stimuli and Individuals

We reviewed empirical evidence highlighting that exogenous and endogenous factors modulate the observed correlates of moral cognition during narrative processing at behavioral and neural levels. Accordingly, we examine moral information across three levels of analysis: 1) stimulus and content; 2) individual differences in traits and behavior; and 3) individual differences in neural patterns. On the first level, our analyses focus on narratives that primarily describe characters who violate particular moral foundations (e.g., care, fairness, loyalty, authority, and sanctity). These semantically similar moral scenarios are embedded in increasingly naturalistic environments across our datasets. On the second level, we examine how individual differences in traits (e.g., empathy, moral intuition salience, political orientation) and moral judgment (e.g., moral wrongness ratings) relate to neural patterns when processing morally relevant scenarios in increasingly contextualized environments. On the neural level, we rely on a controlled, moral judgment task to identify ROIs that are preferentially activated during moral judgment. By doing so, we use the same ROIs across our datasets and analyses to enable cross-dataset comparisons. In bridging all three levels of analysis, we pursue a central question: Which individual differences in traits and moral judgment modulate shared neural responses when processing moral narratives, and how is this similarity affected by the naturalistic complexity of the underlying stimulus?

Answering this question lends itself to intersubject representational similarity analysis (IS-RSA; Finn et al., 2020) which allows us to examine how endogenous between-subject variability in moral traits and judgments (behaviors) relates to similarity in neural responses when processing exogenous morally relevant narratives that are similar in task (stimulus) and content. Specifically, IS-RSA combines two recent developments in neuroimaging analysis: the geometric mapping of relationships between stimulus features (here: moral interactions), as proposed in RSA (Kriegeskorte et al., 2008), and the similarity of computations in a specific brain region across participants, as proposed in intersubject connectivity (Hasson, 2004; Nastase et al., 2019). Thus, IS-RSA allows us to map variations in brain processes evoked by particular moral scenarios onto individual differences in morally-relevant traits and behaviors, effectively testing whether neural activity associated with moral cognition are (dis)similar for participants who are (dis)similar in traits and moral judgment.

Results

We report analyses of three different datasets spanning five stimuli (tasks) that continually increase in naturalistic complexity (Table 1). For each stimulus, we examine the coherence between individual differences in morally-relevant traits and behavior and neural representations as indexed by IS-RSA. Thereafter, we pool the resulting IS-RSA results to assess how brain-behavior-trait coherence is affected by the underlying naturalistic complexity of the stimulus.

Table 1. *Overview of datasets and paradigms*

Stimulus	**Modality**	**Naturalism**	**Moral Features**	**Traits (Behavior)**	***N***
Moral Foundation Vignettes	Visual	Short, static, decontextualized	Harm Cheating Betrayal Subversion Degradation	Pol. Orient., Empathy, Moral Foundation Salience (Moral Wrongness Ratings)	64
Narrated Movie Summaries	Auditory	Short, dynamic, semi-contextualized			
Political Attack Ads	Audiovisual	Short, dynamic, semi-contextualized			
Soap Opera Clips	Audiovisual	Long, dynamic, fully contextualized	Fairness/ Reciprocity Cheating/ Social Sabotage	Empathy, (Moral Wrongness Ratings)	28
Naturalistic Neuroimaging Database	Audiovisual	Long, dynamic, fully contextualized	Harm Cheating Betrayal Subversion Degradation	NIH Toolbox	20

Moral Foundation Vignettes

Background and Procedure

We first assessed where violations of moral foundations – the common currency across our datasets – are encoded in the brain. To this end, we relied on moral foundation vignettes (MFV) (Clifford et al., 2015) to identify regions of interest (ROIs) that encode moral transgressions. In the MFV paradigm, sixty-four healthy subjects underwent functional magnetic resonance imaging (fMRI) while judging the moral wrongness of violations of each moral foundation and social norm transgressions as controls. The MFV feature only sparse information about the identities of perpetrators and targets and do not provide any reason, intention, and consequence for actions. Thus, the MFV offer a controlled paradigm to study where decontextualized moral violations are encoded in the brain.

Behavioral Results

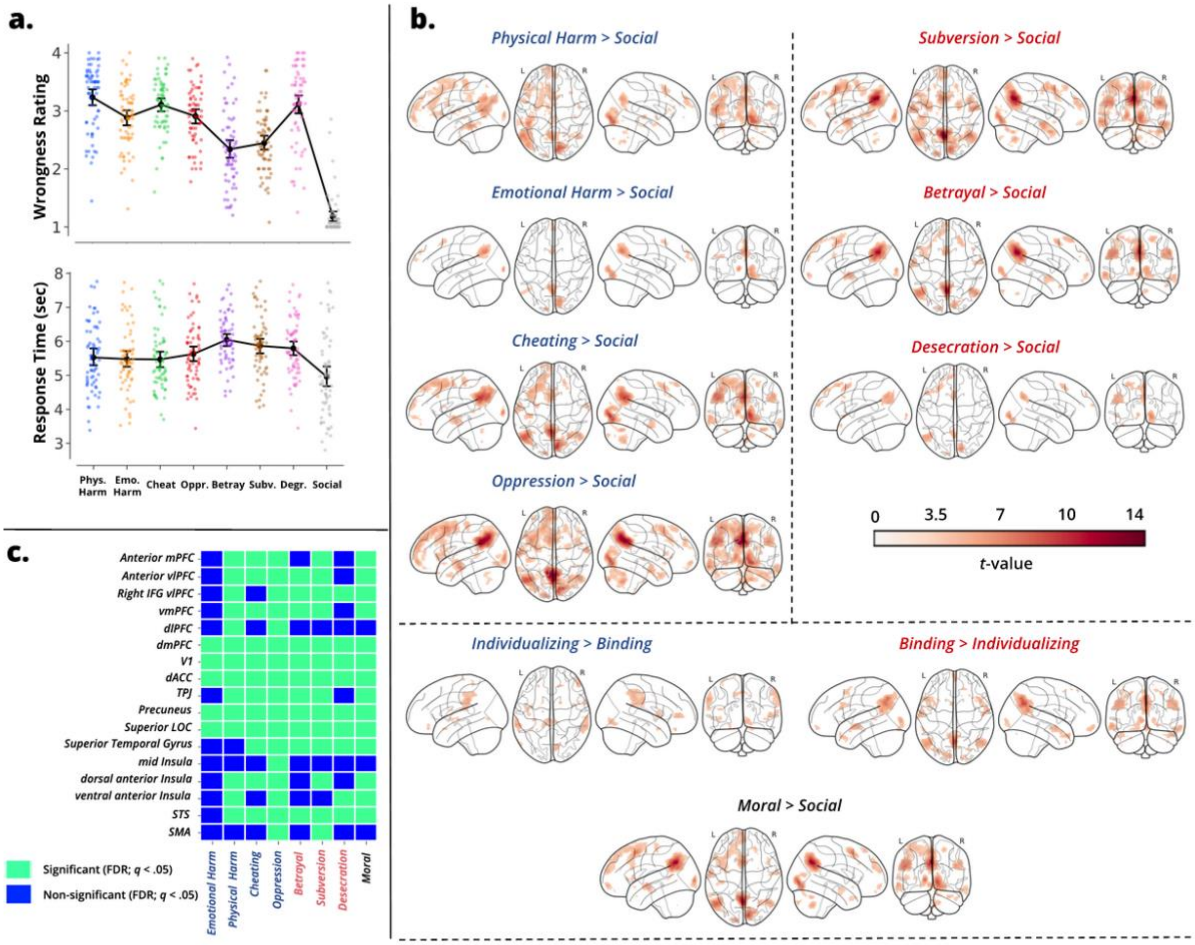
We find that subjects rated all vignettes pertaining to moral, compared to social, norm violations as significantly more morally wrong (Figure 1a, top), suggesting that contrasting subjects' neural responses between moral versus social transgressions indeed captures differences in moral *judgment*, rather than mere *processing* of different transgressions types (Parkinson et al., 2011). Response times to social norm transgressions were also significantly faster than response times to moral transgressions (Figure 1a, bottom), suggesting that evaluations of moral, compared to social, violations elicit a deeper evaluation of individuals' motives and actions and how they relate to one's own values.

Region of Interest (ROI) Analysis

We examined the neural correlates underlying the moral judgment of vignettes with a univariate whole-brain General Linear Model (GLM), contrasting each type of moral violation with social norm violations as controls. Activated regions were identified on the basis of 50 nonoverlapping ROIs derived from a whole-brain parcellation using coactivations from over 10,000 published studies (Vega et al., 2016, see Methods and *Supplemental Information, SI*). This revealed neural activation in a large, dissociable brain network observed in previous work (Figure 1b; Buckholtz & Marois, 2012; FeldmanHall & Mobbs, 2015; Greene & Haidt, 2002; Parkinson et al., 2011; Wasserman et al., 2017; Young & Dungan, 2012).

Notably, each contrast yielded significant ($q < .05$; FDR corrected) activation in the dorsomedial prefrontal cortex (dmPFC), supporting previous research that demonstrated consistent dmPFC activation when judging different kinds of moral transgressions (Parkinson et al., 2011). Dorsal anterior cingulate cortex (dACC), Precuneus (PC), superior lateral occipital complex (LOC), and V1 were also activated across all contrasts; ROIs which also have been associated with moral judgment (FeldmanHall et al., 2014; FeldmanHall & Mobbs, 2015; Young & Saxe, 2008). With the exception of physical care and purity violations, the temporoparietal junction (TPJ), the anterior ventrolateral prefrontal cortex (vlPFC), and the ventromedial prefrontal cortex (vmPFC) also showed significant activation across each moral violation. Physical, but not emotional, harm elicited activation in dorsal and ventral anterior insula, which have previously been linked to perception of bodily harm (Jabbi et al., 2008; Wright et al., 2004).

To directly compare these results with prior research, we generated a brain mask via *Neurosynth* (Yarkoni et al., 2011) consisting of meta-analytic activation maps from 87 studies that contained the word “moral” in their abstract. Applying this mask to a contrast featuring all moral scenarios versus social norms (Figure 1b, bottom), we found a direct overlap with extant moral judgment studies in left dmPFC and PC, right TPJ, and bilateral angular gyrus (Table 2).



	Emotional Harm	Physical Harm	Cheating	Oppression	Betrayal	Subversion	Desecration	Moral
Anterior mPFC	🟦	🟩	🟩	🟩	🟦	🟩	🟦	🟩
Anterior vlPFC	🟦	🟩	🟩	🟩	🟩	🟩	🟦	🟩
Right IFG vlPFC	🟦	🟩	🟦	🟩	🟩	🟩	🟩	🟩
vmPFC	🟦	🟩	🟩	🟩	🟩	🟩	🟦	🟩
dlPFC	🟦	🟩	🟦	🟩	🟦	🟦	🟦	🟦
dmPFC	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
V1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
dACC	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
TPJ	🟦	🟩	🟩	🟩	🟩	🟩	🟦	🟩
Precuneus	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Superior LOC	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Superior Temporal Gyrus	🟦	🟦	🟩	🟩	🟩	🟩	🟩	🟩
mid Insula	🟦	🟦	🟦	🟩	🟦	🟦	🟦	🟦
dorsal anterior Insula	🟦	🟩	🟩	🟩	🟦	🟩	🟦	🟩
ventral anterior Insula	🟦	🟩	🟦	🟩	🟦	🟦	🟩	🟩
STS	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩
SMA	🟦	🟦	🟦	🟩	🟦	🟩	🟦	🟦

Figure 1. Results of univariate GLM analysis for moral foundation vignettes. **a.** Ratings of moral wrongness (top) and response times (bottom) across vignette conditions. Error bars denote 95% confidence intervals. **b.** *t*-maps showing activation unique to each individualizing (blue) and binding (red) moral violation versus social norm. All images are cluster-level corrected (FDR, $q < 0.05$). **c**. ROIs involved in the perception of the 7 types of violations versus control.

We triangulated these encoding models using multivariate pattern analyses to test whether identified ROIs contain more fine-grained information about the categorical divisions and representational organization of moral versus social norm violations. A multivoxel pattern classifier was able to discriminate between moral versus social vignettes in all identified ROIs (*SI*, Figure 1), with particularly high accuracy in posterior cingulate cortex (PCC), anterior vlPFC, and dmPFC. In addition, representational similarity analysis (RSA; Kriegeskorte et al., 2008) revealed a salient distinction between binding and individualizing foundations, as well as between moral versus social norm violations in V1, dmPFC, dlPFC, Precuneus, and dACC (*SI*, Figure 2–3).

Table 2. *Moral vs. social norm violations masked via meta-analytic moral associations*

	MNI Coordinates				
Region name	x	y	z	*t* value	Volume (mm)
L dmPFC	–2	52	38	4.52	2072
L Angular Gyrus	–52	–66	24	7.97	1344
L Precuneus	–2	–58	32	11.25	744
R TPJ	54	–60	24	4.17	256
R Angular Gyrus	48	12	–32	3.36	120

Note. All regions survived cluster-level correction (FDR, q < 0.05).

Taken together, these findings highlight that violations of moral foundations elicit dissociable cortical activation patterns, providing further evidence that morality is better understood via a pluralistic, rather than domain-general framework (Parkinson et al., 2011; Sinnott-Armstrong, 2016; Sinnott-Armstrong & Wheatley, 2012). While dmPFC, PC, V1, dACC, and superior LOC were involved in the judgment of all moral foundations, MVPA demonstrated that the neural representation of each foundation in these ROIs are highly distinguishable (*SI*, Figure 1).

Intersubject Representational Similarity Analysis

We tested how intersubject variability in neural processing of MFV is related to individual differences in traits and behavior using intersubject representational similarity analysis (IS-RSA; Finn et al., 2020; van Baar et al., 2019). The intuition behind IS-RSA is that participants who occupy a similar position in the individual difference space will also be processing information about the MFV more similarly, and regions involved in these processes should show a commensurate similarity in neural activity patterns (Chen et al., 2020; van Baar et al., 2019). We assessed participants' individual differences with regard to traits that have shown reliable associations with moral judgment (Dawson et al., 2021), including multidimensional measures of empathy (Davis, 1980) and moral intuition salience (Graham et al., 2011), as well as political orientation (Eriksson et al., 2019; Graham et al., 2009). In addition, we also computed individuals' similarity for wrongness ratings of each vignette category. One participant was discarded because this person selected the same middle-point option across all empathy items, resulting in a total of 63 remaining subjects.

We generally predicted that individuals who scored highly in a given trait or behavioral domain display a similar brain pattern to other high scorers, whereas low scorers will not look particularly similar to one another or to high scorers. To operationalize this "Anna Karenina" (AnnaK, Finn et al., 2020; see Glossary for detailed description and interpretation) structure of similarity, we computed the mean of every subject pair's rank on a given subscale, and normalized by the highest possible rank. Accordingly, for empathy, we computed four AnnaK models, reflecting individual trait differences in empathic concern (EC), perspective taking (PT), personal distress (PD), and fantasy (FS). In addition, we also computed the itemwise correlation (see Nearest Neighbor item-wise, Chen et al., 2020; Finn et al., 2020) across all empathy items to test whether people who fill out a questionnaire in more similar ways, regardless of their summary trait scores, show similar brain activity. Likewise, for moral intuition salience and wrongness ratings, we computed AnnaK models for each foundation separately to test if individual differences in moral traits and behavior predict neural similarity in a foundation-specific fashion, as well as two NN-itemwise models comprising all moral intuition items and responses to each moral vignette item. For political orientation, we computed both the AnnaK similarity and pairwise Euclidean distance (see Nearest Neighbor, Finn et al., 2020) to test whether individuals of similar political orientation – both in terms of their relative and absolute position on the scale – will show similar brain responses.

We then searched for brain regions that showed a similar representational geometry to these individual similarity measures in terms of the multi-voxel activity pattern correlations between each pair of participants and for each moral vs. social contrast computed by our GLM (Figure 1b). To reduce the search space in the brain while performing this computation, we used the same 50-ROI whole-brain parcellation as before, and we identified ROIs that survived Bonferroni correction (i.e., $p < 0.001$). Following established practices for conducting IS-RSA[1], we examined the overall distribution of IS-RSA values (Spearman's rhos) per individual difference model using a one-sample *t*-test to assess if there is generally some level of representational similarity between brain, traits, and behavior. Note that this does not reveal which, if any, individual nodes show significant representational similarity, but it does show whether there is significant representational similarity at the whole-brain level. In addition, we compared the distribution of Spearman's rho for ROIs preferentially activated by the moral vignettes ("moral ROIs", Figure 1c) to ROIs that did not show significant activation in the vignette task ("control ROIs").

[1] https://naturalistic-data.org/content/Intersubject_RSA.html#comparing-different-models

We observed that each of our individual difference measures was linked to significant intersubject representational similarity (IS-RS) effects at the whole-brain level (*SI*, Figure 3). Overall, the PD dimension of empathy fit the whole-brain neural data best (*t* = –13.33, *p* < .000), with the negative trend of Spearman's rho emphasizing that individuals who are not easily distressed by the experiences of others responded more similarly to other low-scorers, whereas individuals high in PD did not respond similarly to both high and low-scorers. This effect was significantly more pronounced in moral ROIs (*t* = –3.94, *p* < .000). Conversely, individuals high in self-reported PT (*t* = 9.48, *p* < .000), EC (*t* = 8.09, *p* < .000), and FS (*t* = 6.24, *p* < .000) responded more similarly to other high scorers, whereas low-scorers in these domains responded more dissimilarly to both low and high-scorers. Again, for both PT (*t* = 4.64, *p* < .000) and EC (*t* = 3.82, *p* < .000), the effects were significantly higher in moral ROIs.

Similarly, the NN-itemwise empathy model also revealed that individuals who responded to the entire questionnaire in a more similar fashion also processed moral violations more similarly (*t* = 12.56, *p* < .000), again with a higher brain-trait similarity for moral ROIs (*t* = 3.61, *p* < .000). Together, these results corroborate the important, but complex relationship of various dimensions of empathy with moral cognition (Dawson et al., 2021; Decety, 2021).

Moreover, we find that individual differences in political orientation as assessed by the AnnaK model fit the data slightly better than the NN model (compare distributions of RSA values in Figure 2d), emphasizing that individuals' absolute, rather than relative, similarity in political orientation is a better predictor for shared neural representations of abstract moral violations. Surprisingly, the general left-skewed distribution of IS-RSA values for the political orientation AnnaK model across vignette conditions (Figure 2d) does neither support the notion that liberals respond more similarly to the individualizing foundations, nor that conservatives respond more similarly to both individualizing *and* binding foundations. Rather, this finding suggests that liberal-leaning individuals are generally more similar in their representation of moral violations than conservatives. Significant differences between moral and non-moral ROIs only emerged for the political orientation NN model ($t = 2.08$, $p = 0.038$).

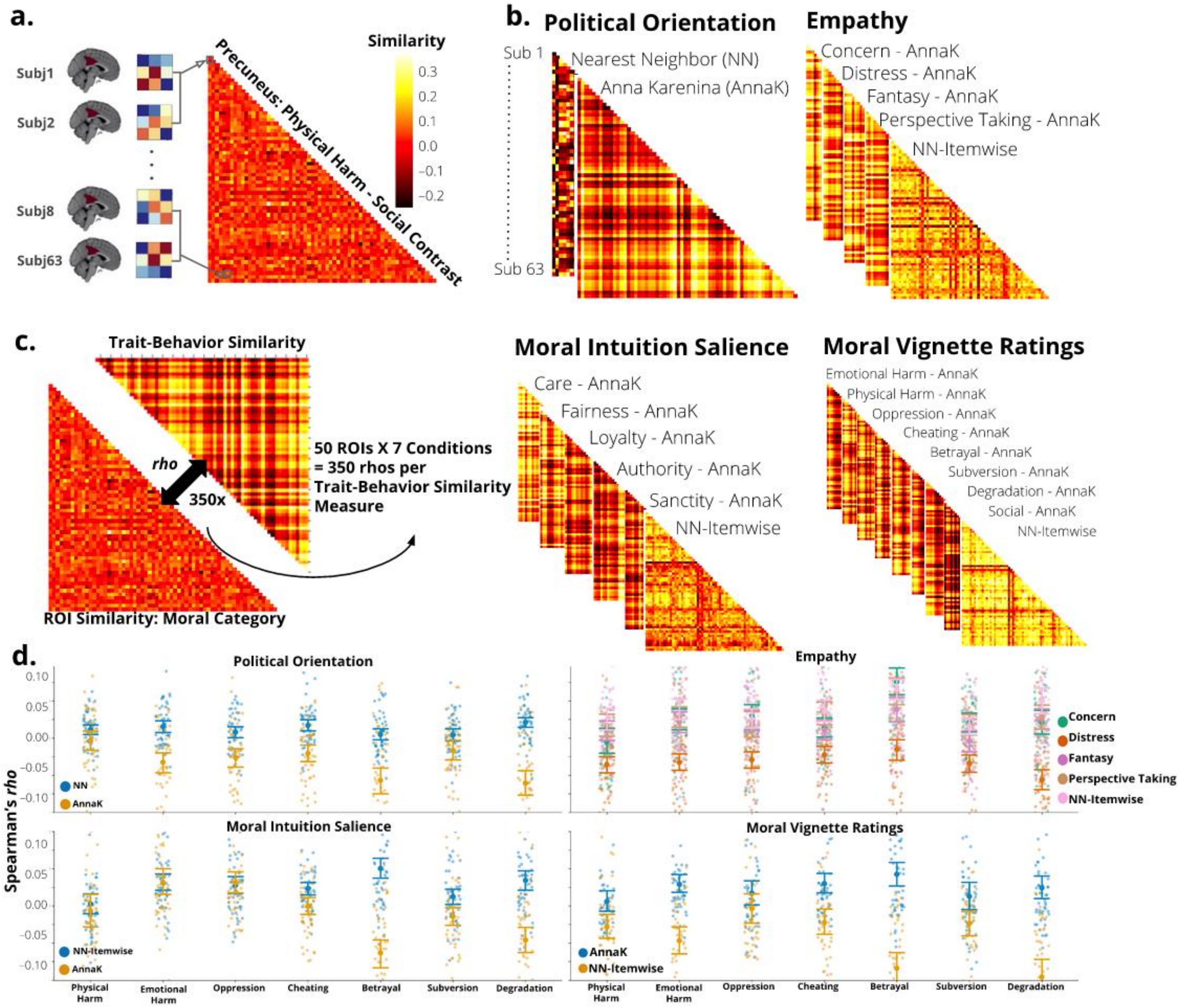


Figure 2. Intersubject representational similarity analysis for moral foundation vignettes. **a.** We first computed the pairwise similarity across each participant's moral vs. social vignette contrast map (one per moral domain, 7 total) in each of the 50 ROIs using 1 – correlation distance. **b.** Next, we computed pairwise individual similarity in political orientation, empathy, moral intuition salience, and moral vignette ratings **c.** We then assessed intersubject representational similarity via Spearman's rho between each ROI, vignette category, and individual difference matrix. **d.** Spearman's rho denotes the strength of brain-behavior-trait coherence across moral vignette categories.

With respect to moral intuition salience, we observed that the NN-itemwise model (*t* = 10.85, *p* < .000) yielded stronger effect sizes than the foundation-specific AnnaK model (*t* = –2.69, *p* = .008), suggesting that individuals' overall similarity in moral intuition salience is a better predictor for neural responses to any class of moral violation than individuals' absolute importance of a foundation that is currently evaluated. In contrast, individuals' absolute moral wrongness ratings for each vignette condition (*t* = –10.55, *p* < .000) yielded a stronger brain-behavior coherence than overall (NN-itemwise) similarity in vignette ratings (*t* = 7.78, *p* < .000), demonstrating that on the behavioral level, similarity in neural patterns to distinct moral violations may be better reflected in individuals' foundation-specific moral wrongness ratings.

Across all IS-RSA analyses, a total of 34 ROIs survived Bonferroni correction (Table 3; for full specification see *SI*, Table 2). With the exception of political orientation, every individual difference matrix was linked to significant neural similarity in at least one ROI. Compellingly, most significant ROIs were identified via the AnnaK moral wrongness ratings, particularly for processing purity violations, providing further evidence that individual differences in dynamic moral behavior, rather than stable traits, may be a better predictor for neural activity patterns related to sparse, controlled, and decontextualized moral stimuli. Moral ROIs that showed significant intersubject representational similarity included, among others, PCC/precuneus relating perspective taking and oppression (*r* = .276, *p* = .001), dmPFC relating NN-itemwise moral intuition salience and subversion (*r* = .13, *p* < .000), dACC relating wrongness ratings and physical harm (*r* = –0.20, *p* < .000), subversion (*r* = –0.23, *p* = .001), and degradation (*r* = –.31, *p* < .000), and mid insula relating wrongness ratings and betrayal (*r* = –0.26, *p* = .001).

Notably, the highest IS-RS was observed in V1 between the AnnaK moral intuition salience models and physical (*r* = .39, *p* < .000) as well as emotional harm (*r* = .32, *p* = .000). This result provides further evidence for the attentional capture of moral stimuli (Gantman et al., 2020; Gantman & Van Bavel, 2014, 2015), especially if they directly relate to human survival (e.g., bodily harm) and are highly valued by certain individuals.

Table 3. *Significant Spearman's rho per individual difference model and vignette condition*

ID Measure	**Model**	**P. Harm**	**E. Harm**	**Cheat**	**Oppr**	**Betr**	**Sub**	**Degrad**	**∑**
Empathy	AnnaK Fantasy	–	–	–	–	–	–	1	1
	NN-Itemwise	–	–	–	–	1	–	1	2
	AnnaK Perspective Taking	–	–	–	1	–	–	–	1
Moral Intuition Salience	AnnaK	1	1	–	–	1	–	–	3
	NN-Itemwise	2	–	–	1	1	1	1	6
Moral Wrongness Ratings	AnnaK	1	4	–	1	2	2	8	18
	NN-Itemwise	1	–	1	–	–	1	–	3
∑		5	5	1	3	5	4	11	**34**

Note. Condition legend: P. Harm: Physical harm; E. Harm: Emotional harm; Cheat: Cheating; Oppr: Oppression; Betr: Betrayal; Sub: Subversion; Degrad: Degradation. Except for the NN-Itemwise model, all reported models were tested with an AnnaK structure.

Narrated Movie Summaries

Background and Procedure

To study how individual differences in moral traits and judgment modulate neural processing of moral stimuli in a more dynamic, but still experimentally controlled environment, the same participants as described previously underwent fMRI while listening to ten professionally narrated movie summaries (for full paradigm description, see *Methods*). In each plot summary (2 per moral foundation), the main character (a) violates one of the five moral foundations while adhering to all other foundations during the "action" part of the story and (b) is either punished or rewarded for their moral transgressions during the "outcome" part of the story. Trial order was randomized and counterbalanced between subjects so that the same order of plots was given to a pair of subjects, but with the *alternative* outcomes. Thus, the action portion of each story featuring a moral violation was listened to by all subjects, albeit in a randomized order. Subsequent to undergoing fMRI, subjects listened to each plot summary again and rated the degree to which they perceived (a) the main character's actions to be moral–immoral, (b) the outcomes that befell the character to be good–bad, and (c) the behavior of the character to uphold–violate each of five moral foundations. One additional participant had to be dropped due to audio issues, producing a final sample of 62 participants.

Treatment Check and Behavioral Results

Behavioral results confirmed that participants perceived the manipulated moral foundation to be most strongly violated in each of the ten stories (*SI*, Figure 4a). Furthermore, all positive-outcome summaries were perceived as having a more positive outcome for the main character compared to the negative-outcome summaries (*SI*, Figure

4b). Character outcome did not modulate perceived strength of foundation violation or rated character morality. Supporting the main theoretical tenets of the MIME (Tamborini et al., 2013), with the exception of two stories, self-reported salience of the moral foundation violated by a character negatively predicted perceptions of character morality (*SI*, Fig. 5). Moreover, we observed that perceived character morality positively predicts story enjoyment (*SI* Figure 6), but an expected interaction where character outcome (reward/punishment) modulates enjoyment of morally appraised characters only emerged in two stories. Taken together, these behavioral results replicate findings from a previous behavioral experiment using the same stimuli (Tamborini et al., 2013).

Intersubject Correlation

We first examined the overall similarity in temporal responses in each region separately for each moral foundation condition using intersubject correlation analysis (ISC), focusing only on the "action" part of a story that featured the violation of a moral foundation. Overall, we observed largely uniform ISC values across conditions, although ISC were slightly lower for subversion summaries compared to stories featuring harm ($t(49) = 3.98$, $p = 0.0002$), cheating ($t(49) = 4.60$, $p < 0.0001$), betrayal ($t(49) = 4.08$, $p = 0.0002$), and degradation ($t(49) = 4.83$, $p < 0.0001$). As expected for auditory stimuli, the highest average ISC values were observed in superior temporal gyrus, primary auditory, and superior temporal sulcus. However, there was also significant stimulus-driven neural synchrony in identified moral ROIs, including TPJ, Precuneus, and dlPFC (Figure 3a–e). Notably, the spatial distribution of ISC values across the whole brain was highly consistent across the five conditions (Figure 3f), suggesting that the overall spatial topography of ISC was highly similar across summaries.

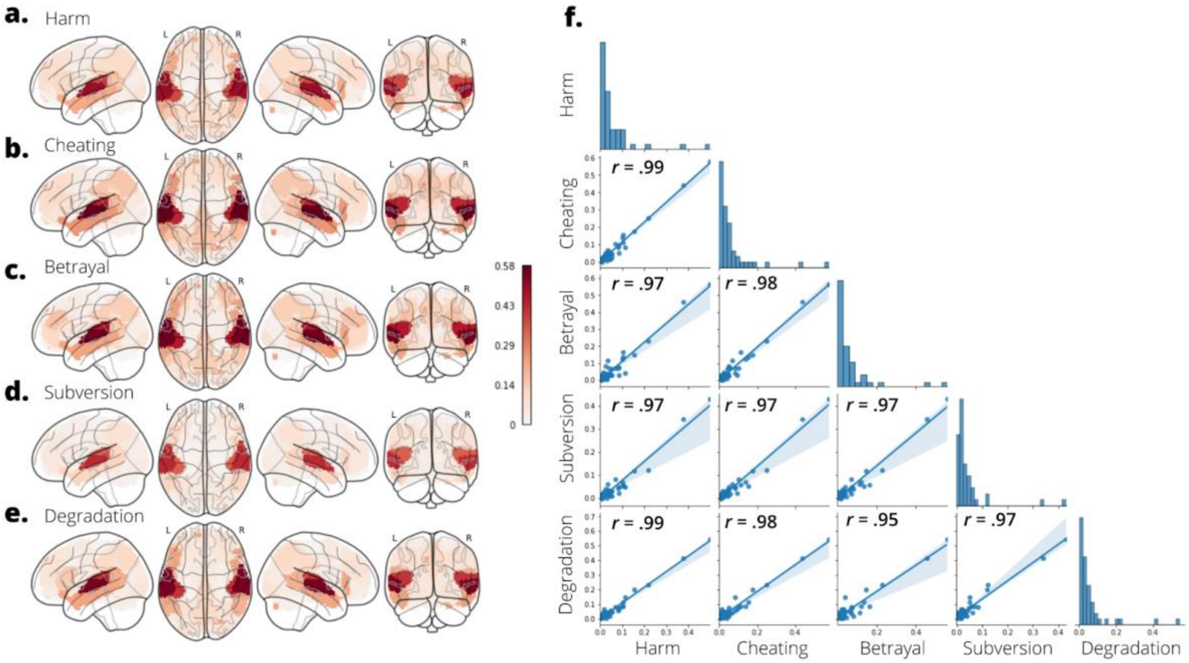


Figure 3. Intersubject correlation for movie plot summaries. Mean ISC within each ROI separately for conditions featuring **a.** harm; **b**. cheating; **c.** betrayal; **d.** subversion; and **e.** degradation. The color bar indicates the average ISC value within each ROI for each condition respectively. **f.** The spatial distribution of these maps was highly similar between the five conditions (each point reflects a single ROI).

Intersubject Representational Similarity Analysis

We again performed IS-RSA to identify brain regions with temporal dynamics that exhibited similar patterns of intersubject variability to individual differences in traits and behavior. We used the same trait-based measures as in the previous IS-RSA, but constructed new behavioral similarity matrices denoting intersubject similarity in perceived foundation transgression and character morality ratings for the movie plot summaries. Because each story pair was judged to mostly violate the transgressed foundation (*SI*, Figure 4a), we computed the average (two per condition) foundation transgression rating for the foundation that was experimentally manipulated in a summary via an AnnaK model as well as an NN-itemwise model that considered similarity in average ratings of *all* foundations for a particular condition. Similarly, we constructed an AnnaK and NN model to capture similarity in perceived character (im)morality.

We find that the moral intuition NN-itemwise model (t = 8.00, p < .000) fit the whole-brain neural data best (*SI*, Figure 8) whereas the moral intuition AnnaK model did not result in significant IS-RS effects at the whole-brain level. Again, this finding demonstrates that relative, overall similarity in moral sensitivities across foundations, rather than absolute weight towards a single foundation that a condition violated, is a better predictor for neural synchrony.

In contrast, we observed that absolute (AnnaK) similarity in rated foundation violation – for the specific foundation that a condition transgressed – yielded stronger effect sizes (t = 5.87, *p* < .000) than overall similarity (NN-itemwise) in rated transgressions of all foundations (*t* = 0.42, *p* = .0675). Replicating findings from the moral vignettes, this result confirms that a foundation-specific IS-RS effect is only prevalent on the behavioral, but not on the trait level when processing moral transgressions. The right-skewed distribution of Spearman's rho also highlights that individuals who perceived a stronger foundation violation were more similar to other individuals with equally high ratings, whereas individuals who did not perceive a strong foundation violation had more idiosyncratic responses. Likewise, the interaction between moral intuition salience and perceived foundation violation ratings were only significant in moral ROIs (*t* = 3.12, *p* = .002). For character morality ratings, we find that the NN model (*t* = –2.65, *p* =.009) is a better predictor for whole-brain shared neural responses than the AnnaK model (*t* = 1.86, *p* = .064). Hence, relative, rather than absolute similarity in rated character morality appears to be a better predictor for neural responses. This effect was significantly more pronounced in moral compared to control ROIs (*t* = 2.59, *p* = .01).

In line with previous research linking individuals' empathic skills and fiction processing (Dodell-Feder & Tamir, 2018; Tamir et al., 2016), we observed that each sub-dimension of empathy yielded significant intersubject RSA values at the whole-brain level. With the exception of distress (*t* = –4.87, *p* < .000), all other dimensions showed a significant positive trend in Spearman's rho, with perspective taking having the strongest effects (*t* = 5.93, *p* < .000), followed by concern (*t* = 5.55, *p* < .000), NN-itemwise, and fantasy (*t* = 4.12, *p* < .000). Observed intersubject RSA values were significantly lower in moral ROIs compared to control ROIs in the distress model, although no such difference was observed for any other sub-dimension of empathy.

Lastly, for political orientation, only the NN model yielded significant intersubject RSA values (*t* = 4.19, *p* < .000), with moral ROIs showing an overall smaller effect compared to control ROIs. Accordingly, for the processing of dynamic, morally-relevant narratives, individuals' relative, rather than absolute similarity in political orientation appears to be a better predictor for shared neural responses.

After Bonferroni correction, three Spearman's rho were statistically significant. The mid insula for the NN-itemwise empathy model in the cheating condition (*r* = .12, *p* = .001), the amygdala for the interaction between moral intuition and perceived foundation ratings in the subversion condition (*r* = .11, *p* < .000), and the right motor cortex for moral intuition in the cheating condition (*r* = –.11, *p* = .001).

Political Attack Advertisements

Background and Procedure

The previously analyzed movie summaries embedded moral violations in dynamic, semi-contextualized environments that more closely resemble everyday environments. Notably, these summaries described fictional characters in an auditory modality without any visual cues that may modulate daily moral cognition. To overcome these limitations, we next analyzed a dataset where the same participants as described previously freely viewed a total of 22 political campaign attack ads (11 ads attacking Trump; “anti-Trump”, 11 ads attacking Clinton, “anti-Clinton”; for full paradigm description, see *Methods*). Each ad embedded information about actors, behaviors, intentions, and consequences within an audiovisual modality. Thus, they offer a more naturalistic, ecologically-valid paradigm to study how individual differences shape moral judgment. Notably, after viewing the ads while undergoing fMRI, subjects re-watched all ads and after each ad answered a battery of questions probing their reaction to the ad, including how (im)moral the candidate’s overall behavior was perceived and whether it adhered to or violated a particular moral foundation.

Behavioral Results

Political affiliation exerted a robust effect on the moral judgment of political candidates (*SI*, Figure 9a). Participants identifying as Democrats perceived Clinton’s behavior in anti-Clinton ads to be more moral compared to Republicans, whereas Republicans rated Trump’s behavior in anti-Trump ads to be more moral compared to Democrats (*SI*, Figure 9b). This effect also persisted for rating the violation of each moral foundation (*SI*, Figure 9c), highlighting that political affiliation reinforces the perception of moral violations by opposing candidates, and dampens the perception that one’s own candidate commits a moral transgression.

Taken together, these behavioral results confirm reported “moral double standards” in political judgment (Eriksson et al., 2019; Voelkel & Brandt, 2019) where differences in moral foundations are primarily driven by ingroup-versus-outgroup categorizations of competing political groups. Independent of political affiliation, individuals perceived Trump and Clinton to violate each moral foundation more than to uphold them across conditions (*SI*, Figure 10). Significant differences between conditions only emerged for the purity foundation, where Trump was rated to violate purity more in anti-Trump ads than Clinton in anti-Clinton ads.

Intersubject Correlation

We again first computed the overall similarity in temporal responses in each region separately for each condition (anti-Trump versus anti-Clinton) using ISC. Overall, we observed largely uniform ISC values across both conditions, although ISC were slightly higher for anti-Trump clips (mean ISC: r = 0.14) than anti-Clinton clips (mean ISC: r = 0.12; t(49) = 5.01, p < .000). As expected for audiovisual stimuli, the highest average ISC values were observed in lower-order regions including superior temporal gyrus, V1, and inferior LOC, but stimulus-driven synchrony also extended to identified moral ROIs including TPJ and STS (Figure 4a–b). The spatial distribution of ISC values across the whole brain was highly consistent across the two conditions (Figure 4c), suggesting that the overall spatial topography of ISC was highly similar across conditions.

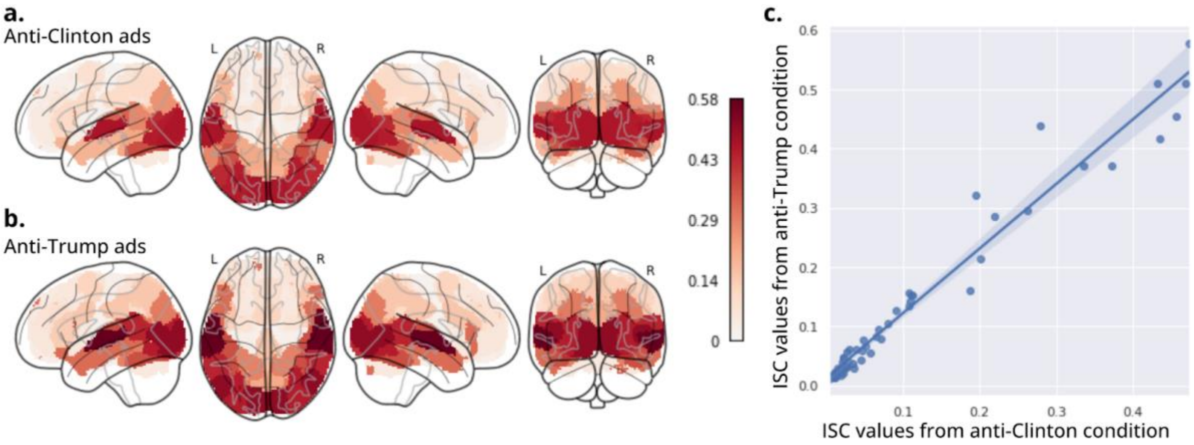


Figure 4. Intersubject correlation for political attack advertisements Mean ISC within each ROI separately for **a.** the anti-Clinton condition and **b.** the anti-Trump condition. The color bar indicates the average ISC value within each ROI for each condition respectively. **c.** The spatial distribution of these maps was highly similar between the two conditions (each point reflects a single ROI).

Intersubject Representational Similarity Analysis

Following our previous framework, we performed IS-RSA to identify brain regions with temporal dynamics that exhibited similar patterns of intersubject variability to individual differences in traits and behavior. We relied on the same trait-based measures as before, and again constructed behavioral similarity matrices denoting intersubject similarity in moral foundation and moral behavior ratings for each political attack advertisement condition. Specifically, we computed NN-itemwise models separately for the anti-Trump and anti-Clinton clips denoting how similarly participants perceived each candidate to uphold or violate moral foundations. Likewise, we created an NN-itemwise model that captured how similarly participants perceived the overall moral behavior – from moral to immoral – of each candidate separately in each condition. For these overall moral behavior ratings, we also constructed an AnnaK model based on the mean morality ratings for each condition. Lastly, for each condition, we created an NN and AnnaK similarity model that captured how favorably participants viewed the respective candidate.

As expected, individual differences in political orientation as captured by an NN model yielded the strongest whole-brain effects (*SI*, Figure 11; $t = 10.01$, $p < .000$), confirming that political orientation is a reliable predictor for explaining shared processing of politically-relevant messages. Compared to the NN model, measuring political similarity via an AnnaK model yielded smaller Spearman's rho ($t = 2.75$, $p = .007$), suggesting that relative, rather than absolute similarity in political orientation is a better predictor for representational similarities during the processing of political attack advertisements. Notably, Spearman's rho for the NN model in the anti-Clinton condition in right vlPFC survived Bonferroni correction ($r = .12$, $p = .0006$), suggesting that within this region, individuals with similar political orientation displayed higher neural synchrony.

With the exception of the concern and fantasy sub-dimensions, Spearman's rho across all empathy models were significantly greater than zero at the whole-brain level, with distress having the highest absolute effect size (*t* = –8.13, *p* = .000), followed by perspective taking (*t* = 7.16, *p* < .000) and the NN-itemwise model (*t* = 5.79, *p* < .000). However, no rho survived Bonferroni correction.

We again find evidence that similarity in moral judgment, rather than moral traits as indexed by moral intuition salience, is a better predictor for explaining shared neural responses. Specifically, participants' relative similarity in rated moral behavior for each candidate yielded the strongest whole-brain coherence with neural processing (*t* = 7.35, *p* < .000), followed by overall similarity in rated moral foundation adherence (*t* = 6.72, *p* <.000). Notably, in the anti-Clinton condition, intersubject representational similarity for rated moral foundation adherence remained significant after Bonferroni correction in the intercalcarine cortex (*r* = .29, *p* = .0002), suggesting that participants' ratings of Clinton's behavior as pertaining to each moral foundation showed commensurate neural activation in this region.

Soap Opera Clips

Background and Procedure

The stimuli in the previously analyzed dataset emphasized the violation of moral foundations and to varying degrees provided contextual information about the identities and intentions of the actors. We now extend this line of inquiry by testing whether observed *consequences* of moral transgressions in the form of punishments and rewards drive intersubject neural response patterns, and whether these similarities in neural responses can subsequently predict individual differences in rated character morality and outcome valence.

Specifically, across full narrative arcs, we examine whether scenarios in which a character is punished for their moral transgressions yield a greater coherence (higher absolute Spearman's *rho*) between neural response patterns across the previously studied ROIs and the moral judgment of characters, compared to scenarios in which a character is rewarded for their moral transgressions.

To answer this question, we use data from a previously conducted study spanning 28 healthy female participants (Weber, 2008). Note that these participants were different from those reported in previous datasets. In this study, participants watched 15 pre-selected scenes from the show *Days of Our Lives*–a popular daytime US network television soap opera–while undergoing fMRI. Notably, these scenes had been selected and analyzed for the presence of moral arcs (who, what, whom, when, why/what intention, and consequence) using a theoretically-driven, methodical content analysis (Weber, 2008; see *Methods*). Based on these content codings, the 15 selected scenes were grouped into the following five experimental groups (3 per group): (1) main character acts morally and experience a positive consequence (moral–reward); (2) immoral behavior and negative outcome (immoral–punishment); (3) moral behavior and negative outcome (moral–punishment); (4) immoral behavior and positive outcome (immoral–reward); (5) neutral behavior (perception of behavior as neither moral nor immoral) and neutral outcome (perception of story outcome as neither negative nor positive). After completing fMRI, participants were seated in front of a PC and watched all 15 scenes again and evaluated their narrative enjoyment of each clip using continuous response measurement. For each viewed clip, participants additionally rated main characters' moral behaviors and outcome valence.

Treatment Check and Behavioral Results

Clip evaluations were as expected (*SI*, Figure 12): the main character was rated as more moral in scenarios where the character behaved more morally, whereas the main character was perceived as more immoral in scenarios that illustrated an immoral character. Outcomes for characters were rated as more positive if the scenario displayed a character that was rewarded, whereas outcomes for characters that were punished were appraised more negatively. Notably, we observed that characters who were punished for their moral actions were perceived as most moral. This result confirms the recently described "virtuous victim" effect, where victims of moral transgressions are attributed greater moral character than neutral controls (Jordan & Kouchaki, 2021). Adding to this result, we show that this effect may be more pronounced if the victim is perceived as highly moral after having committed a moral action and being punished for it.

Intersubject Correlation

We examined the overall similarity in temporal responses in each region separately for each soap opera condition using ISC. While we find largely uniform ISC values across conditions, neural synchrony – particularly in moral ROIs (*SI*, Figure 13) – were higher for scenarios in which an immoral character was punished. Across conditions (Figure 5a–e), the highest average ISC values were observed in superior temporal gyrus, V1, and V2. The spatial distribution of ISC values across the whole brain was again highly consistent across all conditions (Figure 5f), suggesting that the overall spatial topography of ISC was quite similar across conditions.

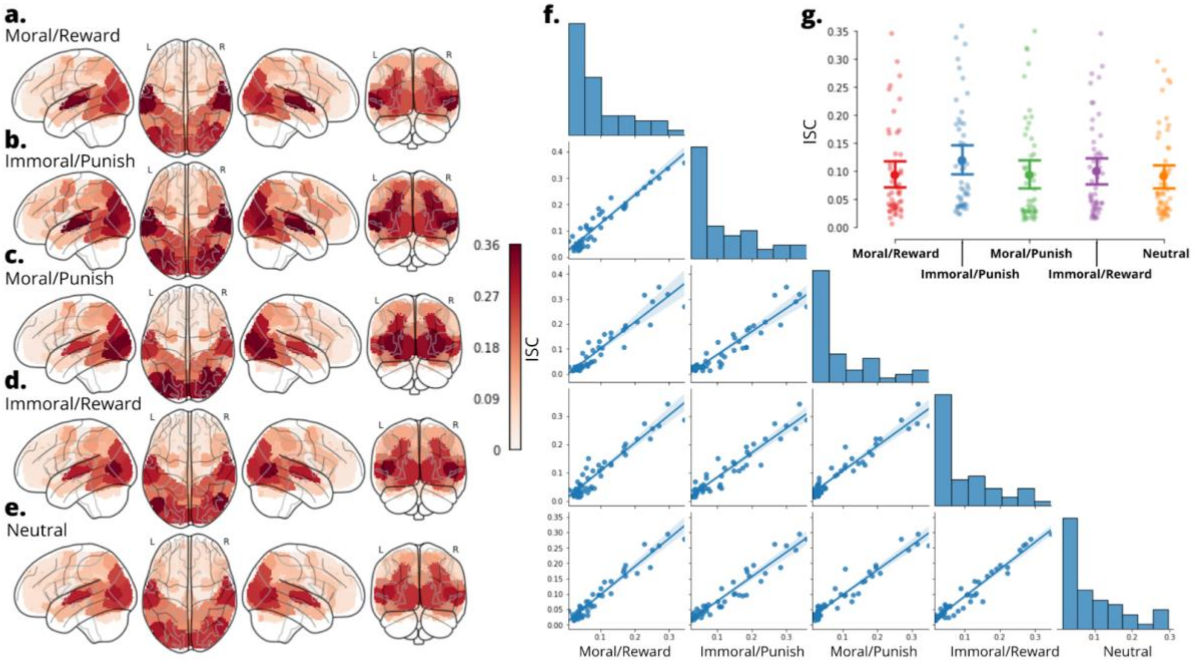


Figure 5. Intersubject correlation for soap opera clips. Mean ISC within each ROI separately for **a.** the moral/reward condition, **b.** the immoral/punish condition, **c.** the moral/punish condition, **d.** the immoral/reward condition, and **e.** the neutral condition. **f.** The spatial distribution of these maps was highly similar between the five conditions (each point reflects a single ROI). **g.** Mean ISC was highest in the immoral/punish condition.

Intersubject Representational Similarity Analysis

To relate individual differences in traits and moral judgment to neural similarity in processing soap opera clips, we performed IS-RSA. Because subjects in this dataset were different from those reported previously, individual difference matrices had to be re-computed. As in the previous dataset, we constructed intersubject similarity matrices denoting trait-level and behavioral-level individual differences. On the trait level, we computed participants' overall similarity in empathy via an NN-itemwise model and again used an AnnaK model for each empathy subdimension spanning empathic concern (EC), perspective taking (PT), personal distress (PD), and fantasy (FS) (Davis, 1987). Likewise, we computed two NN-itemwise models that captured similarity in personality as measured by the Ten Item Personality Index (TIPI; Gosling et al., 2003) and Eysenck questionnaire

(Eysenck & Eysenck, 1993). On the behavioral level, we computed intersubject similarity in moral judgment for character morality and outcome valence using an AnnaK model on the average ratings per condition (three ratings per condition) and an NN-itemwise model capturing the correlation of ratings across clips per condition. Lastly, for each condition, we created an NN-itemwise similarity matrix denoting the pairwise, intersubject correlation for continuous response measurements (CRMs) across the three condition-specific, concatenated clips.

Examining intersubject representational similarity at the whole-brain level, we find the strongest behavior-trait coherence within the AnnaK fantasy subdimension of empathy (t = 13.77, p < .001), providing evidence that individuals who more readily identify and mentalize with characters indeed process audiovisual narratives in more similar ways. For this model, mPFC survived Bonferroni correction in the condition where a good character was rewarded (r = 0.345, p < .001). We also observed that individuals with lower tendency of distress (t = –3.92, p < .001) and perspective taking (t = –4.90, p < .001) process soap opera clips in a more similar fashion compared to their high-scoring counterparts. On the personality level, itemwise response similarity to the TIPI yielded slightly higher (t = –6.85, p < .001) intersubject representational similarities in absolute terms than itemwise responses to the Eysenck questionnaire (t = 2.23, p = .027), whereas in both models, the Spearman's rhos were significantly higher for moral than control ROIs (TIPI moral vs. control: t = 4.93, p < .001; Eysenck moral vs. control: t = 3.25, p = .001).

On the behavioral level, we observed that IS-RS for individual similarities in CRM yielded largely negative Spearman's rho (t = – 5.32, p < .001). This is surprising as it suggests that subjects who were more similar in their moment-to-moment ratings had a *dissimilar* neural activation while processing the soap opera clips. Although speculative, we think this result may reflect that CRMs, although measured dynamically, may lag behind or lead neural activations. Thus, implementing different lags and leads in CRMs may yield a closer match between brain and behavior as assessed via IS-RSA. In contrast, for post-hoc character morality ratings, we indeed observe significant positive IS-RS effects for the NN (*t* = 3.71, *p* < .001) and NN-itemwise (*t* = 2.03, *p* = .044) model, suggesting that participants who judged characters' moral behavior more similarly indeed processed the narrative scenarios in a more similar manner. Particularly, in scenarios where an immoral character was rewarded for their actions, we find that individuals who rated this character as more *moral* showed significantly similar response patterns in Superior Temporal Gyrus (*r* = 0.334, *p* = .001). In addition, when comparing the different soap opera scenarios (*SI* Figure 16), we find that the AnnaK model for character morality yielded the most negative whole-brain intersubject representational similarity in scenarios where a character behaves immorally, particularly in visual attention (V1, *r* = –0.42, *p* = 0.002, V2, *r* = –0.37, *p* = 0.008) and auditory networks (Primary Auditory, *r* = –0.42, *p* = 0.002; Superior Temporal Gyrus, *r* = –0.41, *p* = 0.003). This finding compellingly demonstrates that individuals who processed immoral scenarios in a more similar fashion indeed judged the observed character as more immoral compared to individuals who rated the character as less immoral, providing further evidence for the brain-behavior linkage between attentional networks and moral judgment.

Full Length Movie

Background and Procedure

For the third dataset, we used a full-length movie to assess brain-trait coherence when processing dynamic, morally-relevant scenarios. Movies are emerging as the "goldstandard" for naturalistic neuroimaging (Chang et al., 2021; Chen et al., 2020; Finn et al., 2018; Finn & Bandettini, 2021), and situate moral behaviors in contextualized environments that unfold over long time windows (Hopp, Fisher, & Weber, 2020). Hence, movies provide a high degree of naturalism for assessing how observed moral actions link intersubject variability in traits and behavior. We utilized the recently introduced *naturalistic neuroimaging database* (NNDb v.1.0; Aliko et al., 2020) and obtained data for the movie *500 Days Of Summer*, which was viewed by 20 participants while undergoing fMRI. In addition, we recruited four independent undergraduate research assistants (RA) to obtain continuous response measurements (CRM) of how engaging they perceived each moment of the movie to be. This CRM served as explorative, behavioral assessment to illuminate linkages among a movie's moral content, its behavioral rating, and neural processing. To extract the latent moral information of the movie, we used several techniques from natural language processing: First, we retrieved and computationally parsed the movie's screenplay, providing us the scene and stage directions, action descriptions, characters, and dialogue. Two additional RAs then jointly viewed the movie while labeling the timestamps corresponding to the onsets of each screenplay scene. Additional scenes that were not present in the screenplay were also recorded, while scenes that were missing in the movie were discarded, and scenes spanning multiple scenes in the screenplay grouped together. In total, 97 scenes (*M* duration= 56.16sec; *SD* duration = 61.51sec) were identified.

Tracking Dynamic States of Engagement During Movie Watching

We first examined how behavioral engagement as assessed by the CRM changes over time as individuals comprehend the movie and whether changes are synchronized across participants. Providing initial validation for our self-report task as a measure of stimulus-related narrative engagement, with the exception of RA2, we find that the raters experienced the interestingness of the movie in highly similar ways, with particularly strong correlations between RA1 and RA3 (r = 0.86, p < .000), RA1 and RA4 (r = 0.55, p < .000), and RA3 and RA4 (r = 0.53, p < .000) (Figure 6b–c).

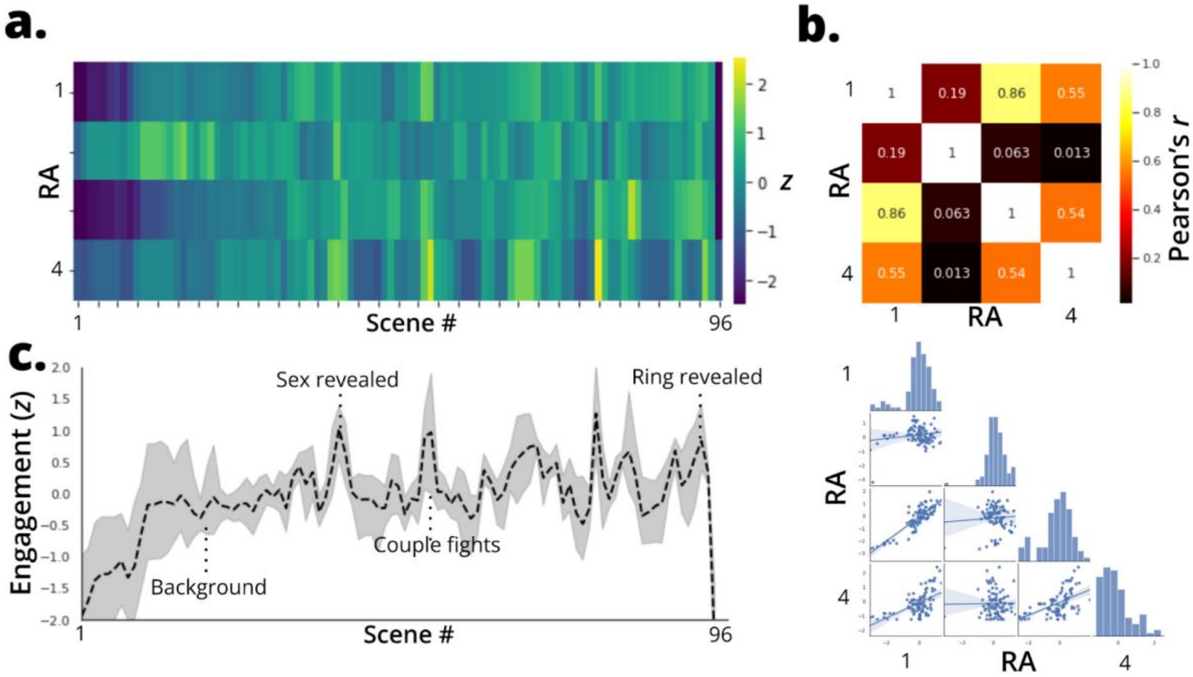


Figure 6. Continuous measurement of engagement while watching *500 Days of Summer*. **a.** Every RA's engagement ratings across scenes. Ratings were z-normalized across scenes for each RA. **b.** Pairwise participants' response similarities. Engagement rating similarity was calculated by Pearson's correlation between pairs of participants' engagement ratings across time. **c.** Average engagement ratings, which are used as proxies for group-level states of engagement. The gray area indicates 95% confidence interval (CI). Event descriptions are given at moments of peak engagement.

Since engagement ratings were similar across participants, we treated the group-average engagement rating (Figure 6c) as a proxy for stimulus-related engagement, common across individuals. We qualitatively assessed moments when participants were, on average, most or least engaged in the narratives (Figure 6c). Engagement peaked at moments when the first intimate moment between the main couple was revealed; when the couple fought over their relationship status; and when the main male character unexpectedly found out that his date got engaged. On the other hand, engagement was generally low when a story setting was being developed or when a protagonist was having an internal thought. The group-averaged engagement time courses were convolved with the hemodynamic response function (HRF) to be applied to a separate pool of individuals who participated in the fMRI study.

Engagement Ebbs and Flows with Moral Content

Given that RA's engagement dynamics were time locked to the stimuli, we examined whether morally-relevant features of the narrative drove changes in engagement. Specifically, we assessed the degree to which words in the time locked subtitles of the movie signal upheld, violated, or conflicted (i.e., violating moral norms in order to uphold others (Tamborini, 2011, 2013) moral foundations. To this end, we used the eMFDscore software (Hopp et al., 2021) on the parsed subtitles to calculate the probability that any given word reflects the adherence or transgression of moral foundations. Moral conflict scores were calculated by following procedures developed by Hopp and colleagues (2020). Next, we averaged the moral content features for each scene. We then conducted partial correlations between group-average engagement and moral content features.

We find that scenes featuring moral conflicts between authority-subversion (partial r = 0.23) and sanctity-degradation (partial r = 0.20), as well as notions of betrayal (partial r = 0.16) were most positively correlated with changes in engagement when controlling for the rest of the variables. In contrast, scenes that displayed a moral conflict between loyalty-betrayal (partial r = –0.198), subversion (partial r = –0.182), and harm (partial r = –0.179) were most negatively related to engagement. Because the CRM is only based on four independent raters, we did not conduct any inferential statistics but rather interpret these results as preliminary evidence that engagement ebbs and flows with certain facets of morally-relevant narrative content.

Neural Synchrony of Moral Networks Increases During Engaging Moments of the Story

Next, we examined the overall similarity in temporal responses in each of the 50 parcellated regions separately for each scene using ISC. We find highest mean ISC values in visual and auditory networks (Figure 7a), including inferior LOC (mean ISC = 0.44), STG (mean ISC = 0.40), Lateral Occipital/Temporal Occipital (V2) (mean ISC = 0.39), V1 (mean ISC = 0.37) and primary auditory (mean ISC = 0.35). We also observe notable mean ISC in many of our previously identified moral ROIs, including TPJ (mean ISC = 0.26), STS (mean ISC = 0.2), Cerebellum (mean ISC = 0.18), PCC/Superior LOC (mean ISC = 0.18), and PCC/Precuneus (mean ISC = 0.17), suggesting that participants also had some shared interpretations of the movie scenes in ROIs implicated in controlled moral judgment. Moreover, mean ISC of individual scenes fluctuated markedly throughout the movie (Figure 7b), indicating that scenes elicited varying degrees of intersubject synchrony.

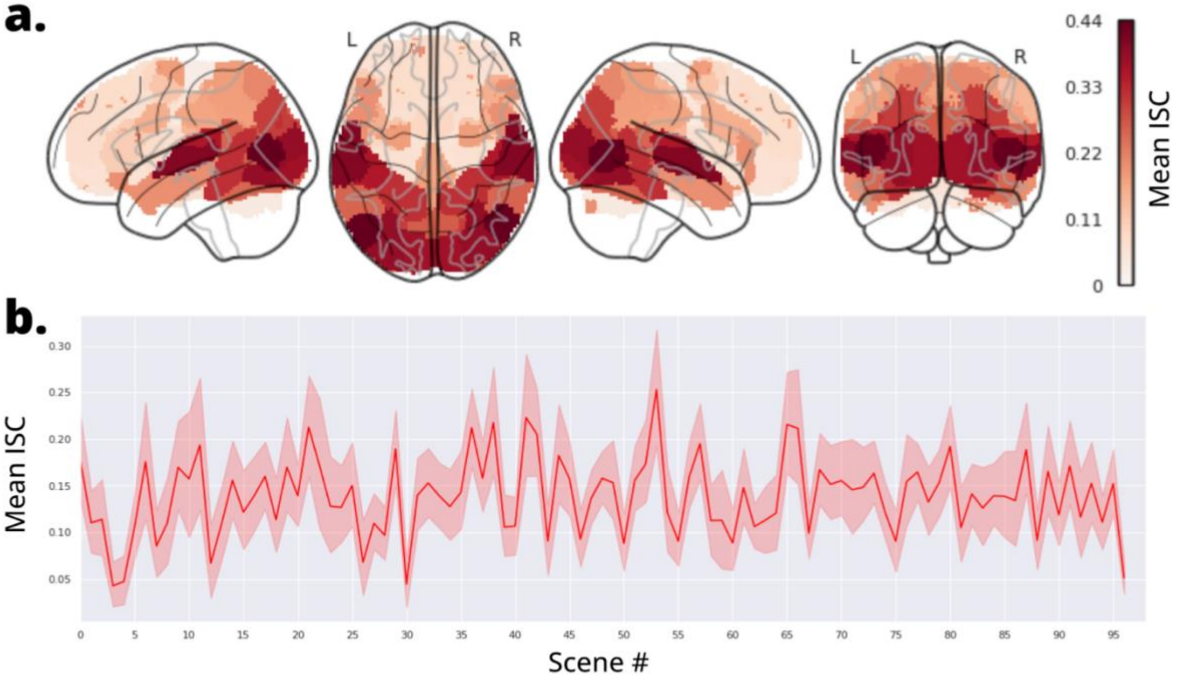


Figure 7. Intersubject correlation for scenes in *500 Days of Summer*. **a.** Mean scene-based intersubject synchrony was highest in visual and auditory networks. **b**. Mean ISC fluctuated markedly across scenes. Shaded areas of the line plot reflect 95% CI based on 1,000 bootstraps.

We then asked whether ISC would increase as participants, on average, become more engaged in the story. We used HRF convolved group-average engagement ratings from the CRM as our index of narrative engagement because participant specific measures of engagement were not available in the existing fMRI dataset. We then correlated the scene-based ISC with the scene-based CRMs for each of the 50 ROIs separately (Figure 8).

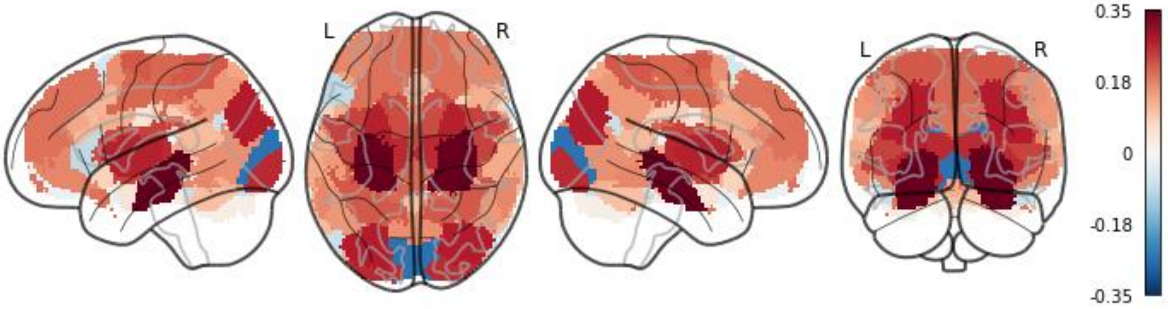


Figure 8. Dynamic ISC scales with changes in engagement. Similarity as indexed by Pearson's correlation coefficient (uncorrected) between scene-based ISC and scene-based engagement ratings.

We observed the highest positive coherence between ISC and engagement ratings in Dorsal Caudate (*r* = 0.353), Superior LOC (*r* = 0.278), Dorsal Anterior Insula (*r* = 0.27), dlPFC (*r* = 0.263), and dmPFC (*r* = 0.192). Surprisingly, the correlation between ISC and engagement ratings was most negative in V1 (*r* = –0.26), suggesting that shared neural activation in the early visual cortex decreased as subjects became more engaged in the narrative. Taken together, these results suggest that BOLD responses are more entrained to the stimulus when self-reported levels of engagement increase, especially in regions implicated in moral judgment (Dorsal Anterior Insula, dlPFC, dmPFC).

Moral Content Modulates Intersubject Representational Similarities

In line with our central research question, we examined whether moral features of a narrative modulate the brain-trait signal during movie viewing. To this end, we first computed the (dis)similarity of the 20 participants viewing the movie in the scanner with regards to a range of emotional measures taken from the National Institute of Health (NIH)

Toolbox (Figure 9; *SI* Figure 17). Thereafter, for each scene, we computed IS-RSA between each individual difference model and ISC of each of the 50 ROIs.

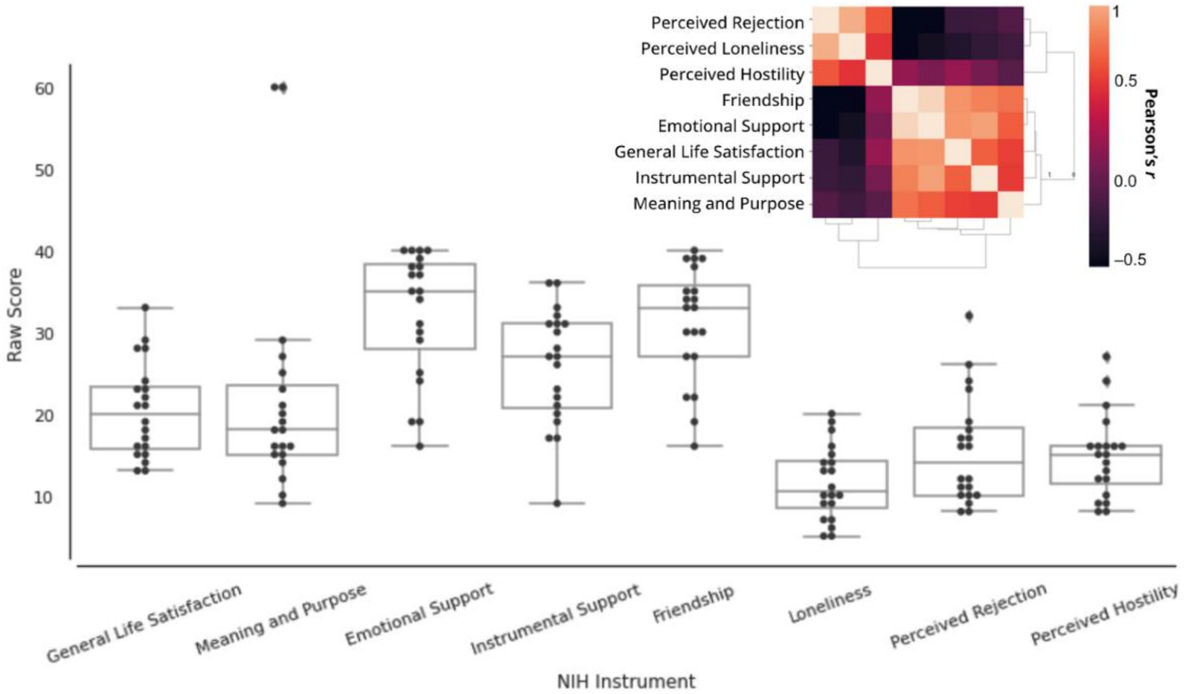


Figure 9. Boxplots and factor structure of NIH toolbox scores for emotional measures. Clusters for the correlation map were generated via the average linkage method.

We find that whole-brain coherence between traits and neural representations varies dynamically throughout the movie, and that individual difference models yield idiosyncratic peaks and troughs during different time windows (scenes) of the narrative (Figure 10a). Hence, future work may apply reverse correlation (Hasson, 2004) approaches to qualitatively study which narrative elements triggered these boosts in Spearman's rhos. Notably, the AnnaK models consistently outperform the NN models in terms of absolute effect sizes (Spearman's rho), suggesting that individuals' absolute position in the individual difference trait space is a better predictor for shared neural processing than their relative distance to each other.

Moreover, significant, scene-based Spearman's rhos (FDR corrected, *q* < .05) across individual difference models were observed most frequently in TPJ and Cerebellum (Figure

10b), confirming previous work that has linked neural similarities in these ROIs to different individual differences measures (Chen et al., 2019; Finn et al., 2018, 2020).

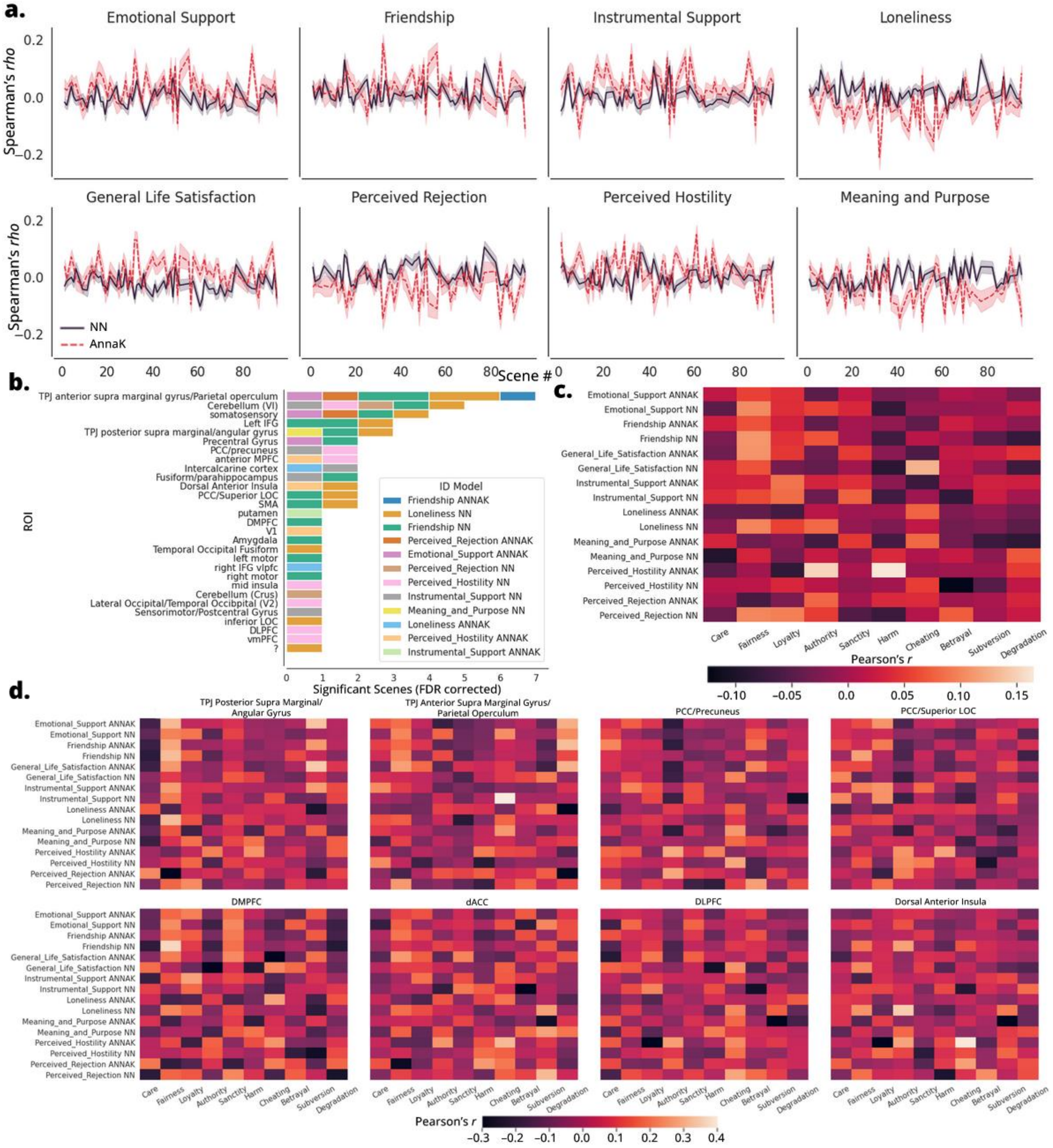


Figure 10. Moral content shapes intersubject representational similarity during movie viewing. **a.** Dynamic, mean whole-brain intersubject representational similarities for individual difference measures. **b.** Significant Spearman's rho were detected most frequently in TPJ and Cerebellum. **c.** Whole-brain coherence between individual difference measures and neural representations of the movie scales with a scene's moral content. **d.** Pearson's correlation coefficient relating moral content and IS-RSA in moral judgment parcels.

Next, we tested whether intersubject representational similarities scale with the moral content of a scene. To this end, for each scene, we correlated the computed moral foundation scores with the Spearman's rho for each of the 50 ROIs and individual difference models (Figure 10c). We find that the brain-trait coherence for the perceived hostility AnnaK model increases in scenes that feature notions of harm (*r* = 0.16) and authority (*r* = 0.15), suggesting that the extent to which individuals perceive their daily social environment as hostile becomes a more relevant predictor of shared neural patterns when processing story elements that feature physical or emotional harms or concerns about maintaining social order. Exploratory analyses (SI Figure 18) showed that content relating to harm in particular positively increases IS-RSA in attentional networks, including Primary Auditory (*r* = 0.450), Superior Temporal Gyrus (*r* = 0.343), and V1 (*r* = 0.35).

Notably, the relatively higher correlation for primary auditory likely reflects the fact that we extracted moral content from subtitles, which primarily deliver auditory (e.g., dialogue) information. This compelling result suggests that individuals who are under higher perceived threat also monitor and respond to perceived harms in narratives in a more similar fashion compared to those who typically navigate through less hostile environments. Moreover, displays of fairness increased the coherence between trait measures of social support (emotional support NN, *r* = 0.102; friendship NN, *r* = 0.108; loneliness NN, *r* = 0.097) and general life satisfaction (*r* = 0.107), highlighting that individuals' perceptions of the availability of friends or companions in daily life becomes a stronger predictor for shared neural representations during moments of the narrative that involve notions of rights and justice.

Lastly, we examined whether these effects are specific to ROIs involved in moral judgment (Figure 10d). We find a robust association between intersubject similarity in social support and companionship within TPJ/angular gyrus during the processing of narrative moments that emphasize fairness, whereas for the same traits, notions of degradation increased Spearman’s rhos in TPJ/Parietal operculum. In particular, displays of cheating correlated markably (*r* = 0.41) with the brain-trait signal in TPJ/Parietal operculum for the instrumental support NN model. Moreover, we observe that the brain-trait coherence for the Friendship NN model increases in dlPFC during moments that emphasize fairness (*r* = 0.36).

Taken together, we find support that moral content in a narrative impacts viewers’ engagement and temporarily increases the brain-trait coherence in ROIs previously linked to moral judgment. Although the obtained individual difference measures were not related to morality *per se* – for example, no measures of empathy or moral intuition salience were collected in this dataset – we were still able to show that narrative features relating to harm increase intersubject representational similarity effects for trait measures of perceived hostility, whereas notion of fairness were processed more similarly by individuals high in social support. Overall, these findings warrant further research into the power of morally relevant movie content to emphasize individual differences in traits and behavior that extend beyond moral traits and moral judgment.

How does naturalistic complexity affect brain-behavior-trait coherence?

In order to examine how the brain-behavior-trait coherence is affected by the underlying naturalistic complexity of a paradigm, we pooled together the intersubject synchrony (ISC; Pearson’s *r*) as well as intersubject representational similarity (IS-RSA; Spearman’s rho) results from our previously analyzed datasets.

Neural Synchrony in Moral Brain Networks Scales with Naturalistic Complexity

We first examined how intersubject synchrony varies as a function of the underlying stimulus (task). Notably, we find that mean ISC in moral brain networks – particularly in TPJ, STS, and Precuneus – increases continuously across vignettes, narrated movie summaries, and political attack advertisements (Figure 11). As the fMRI scanner settings and subjects were held constant across these tasks, this result provides compelling evidence that processing increasingly contextualized moral violations indeed leads to more synchronized neural activity, especially in ROIs that reliably encoded moral information in our previous analyses. As expected, we also observed that ISC in auditory and visual networks scale with the corresponding stimulus modality: for stimuli that contained auditory signals (i.e., narrated movie summaries and political attack advertisements), we observed higher ISC in auditory networks, whereas for stimuli that contained visual information (i.e., vignettes and political attack advertisements), we find higher ISC in visual networks.

Moreover, we observed that processing soap opera clips and full length movies resulted in lower ISC in TPJ and STS compared to the processing of political attack advertisements. This suggests that in political attack advertisements, moral transgressions of political candidates, despite being embedded in a dynamically unfolding attack advertisement, may still become salient to a degree that elicits high ISC in moral brain networks. In contrast, moral cues that are situated in a fully contextualized narrative world may be overshadowed by the numerous additional – and potentially noisy – information that the narrative mode presents, hence resulting in lower ISC in morally relevant brain networks. Yet, as the soap opera clips and the movie dataset contained data with different subjects and fMRI scanner settings, it cannot be ruled out that differences in ISC in these datasets were driven by subject composition or scanning protocols. However, we observed that ISC in ROIs that are not part of auditory, visual, and moral judgment networks (Figure 11d) did not follow any trends observed in these networks across the datasets. This result highlights that it is unlikely that cross-study differences in ISC are primarily driven by differences in sample and fMRI recording devices.

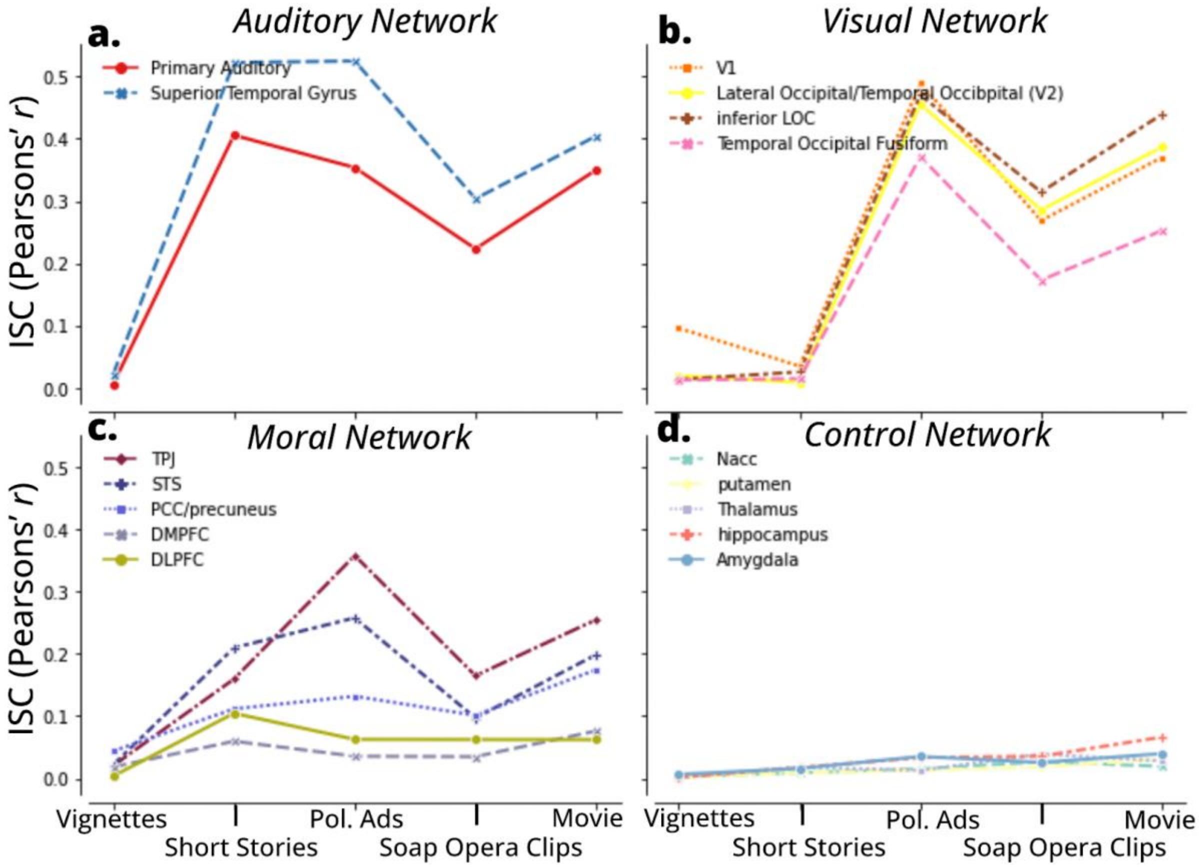


Figure 11. Mean intersubject synchrony across tasks, datasets, and brain networks. Line charts denote mean ISC values averaged over the conditions of each task.

Brain-Behavior-Trait Coherence Across Stimuli of Increasing Naturalistic Complexity

Next, we examined how intersubject representational similarities – indexed by Spearman's rho – with respect to different traits and behavioral measures fluctuate across tasks and datasets. As the movie dataset did not contain any behavior or trait measures that overlapped with our previous datasets, we discarded the IS-RSA results for the movie dataset. Accordingly, we first contrasted the IS-RSA for each trait dimension of empathy across auditory, visual, moral, and our defined control networks (Figure 12). Overall, we do not find support that IS-RSA in moral judgment ROIs increases continually from controlled, decontextualized stimuli to more naturalistic, contextualized narratives. Rather, IS-RSA are largely stable and more nuanced with regard to specific dimensions of empathy. A general positive trend in IS-RSA across datasets is only apparent in TPJ and auditory networks for the fantasy dimension (Figure 12d). This result suggests that individuals who more easily become transported in narratives and empathize with characters indeed become more similar to other individuals high in fantasy (and dissimilar to individuals low in fantasy) in their neural representational patterns as stimuli integrate more narrative features (e.g., character identities, intentions, and consequences).

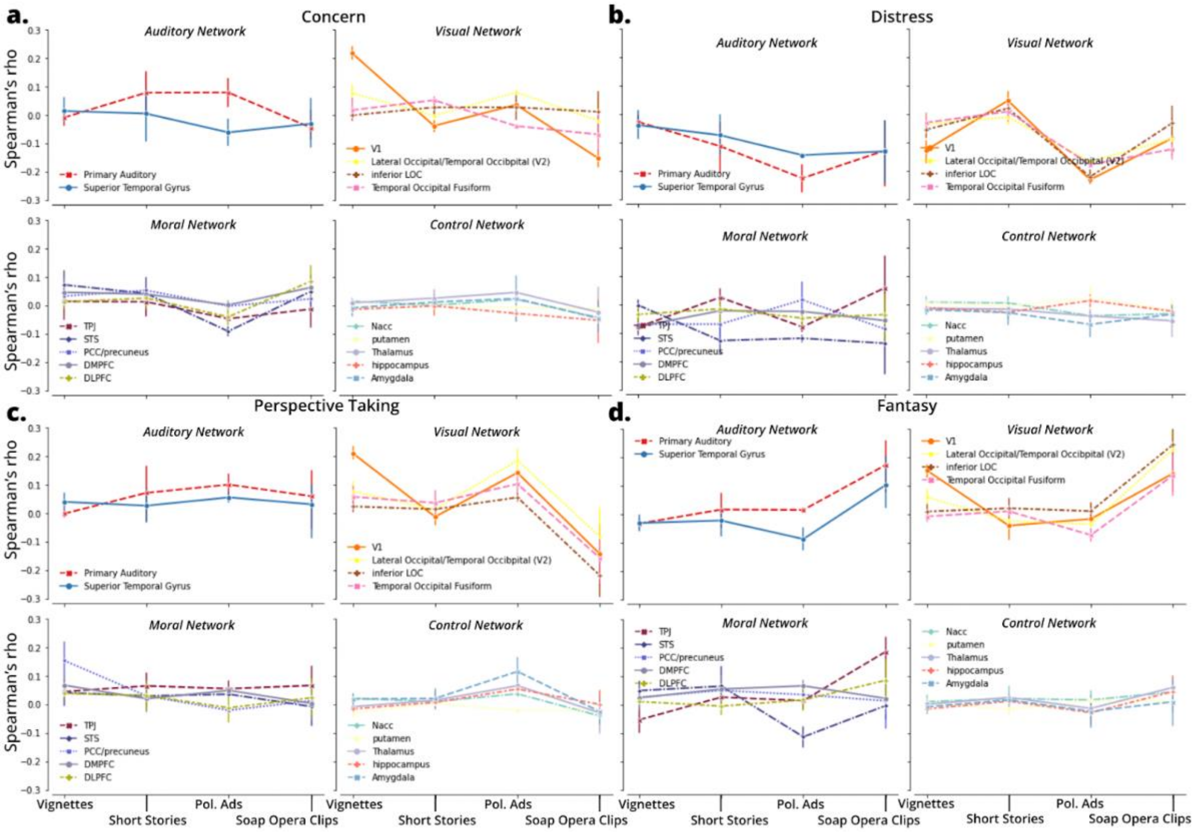


Figure 12. Intersubject representational similarity for empathy across datasets. Intersubject representational similarities (Spearman's rho) for concern (a), distress (b), perspective taking (c), and fantasy (d) across datasets and brain networks. Individual differences for each empathy trait were captured via an Anna Karenina model.

In visual ROIs, we find a negative trend for concern and perspective taking across datasets, indicating that visual activation patterns of individuals who scored low in these dimensions become more similar to other low scorers as the naturalistic complexity of a paradigm increases.

These observed trends largely extend to comparative IS-RSA models on behavioral moral wrongness ratings. Here, for the NN model, we find that Spearman’s rho in moral judgment ROIs increases from short stories to political attack advertisements (Figure 13a), whereas we again observed that the brain-behavior signal was highest for controlled moral vignettes in visual networks. Conversely, for the AnnaK moral wrongness model, a clear negative trend in Spearman’s rho is apparent in primary auditory, V1, as well as in TPJ (Figure 13b).

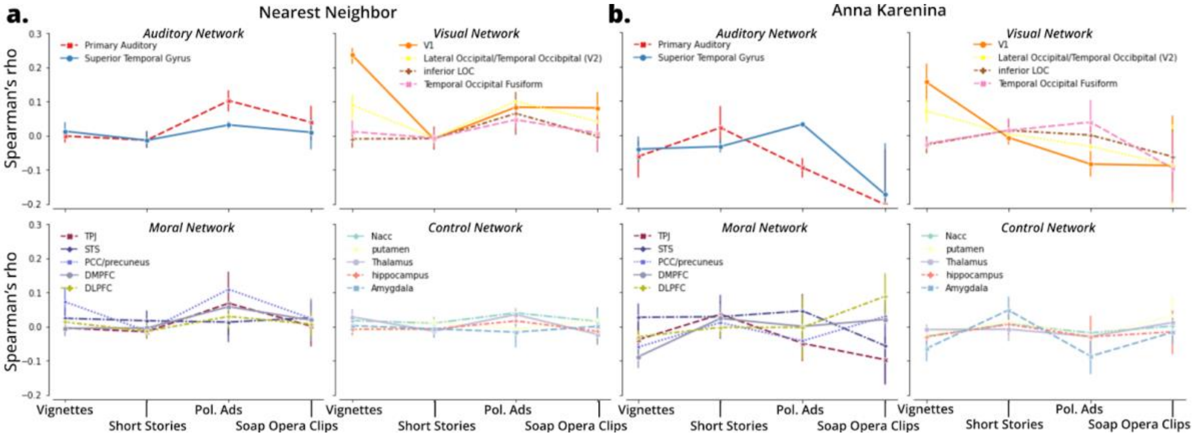


Figure 13. Intersubject representational similarities for moral judgment across datasets. Intersubject representational similarities (Spearman’s rho) for moral wrongness ratings as indexed by a Nearest Neighbor **a.** or Anna Karenina **b.** model.

Due to the signed nature of the AnnaK model, this finding suggests that in decontextualized environments, people that judged behaviors as *moral* had a more similar neural response, whereas during the processing of contextualized narratives, individuals who rated characters as more *immoral* showed commensurate neural activation patterns. Interestingly, this effect diverges in dlPFC and TPJ (compare Spearman’s rhos between political attack ads and soap opera clips): in dlPFC, individuals that judged characters as more moral had a more similar neural response to other high scorers, whereas in TPJ, individuals with more immoral ratings had a more similar response to other individuals who rated characters as more immoral; an effect that was not as pronounced during the processing of other stimuli types.

Taken together, these cross-dataset results do not confirm a straightforward hypothesis linking greater naturalistic complexity to a domain-general increase in observed brain-behavior-trait coherence. Rather, our findings paint a more complex picture, highlighting unique associations between stimulus (task) type, individual difference measure, and intersubject representations in particular brain networks when studying idiosyncrasies in moral cognition.

Discussion

Moral judgments in daily life are shaped by our personal traits and the contextual cues embedded in our dynamic environment. While previous research has largely used highly controlled experimental paradigms to study the neural basis of moral judgment, we herein

examined the challenges and opportunities that naturalistic stimuli offer for advancing our understanding of how individual differences shape moral cognitions above and beyond traditional moral judgment tasks. To this end, we analyzed three different functional magnetic resonance imaging (fMRI) datasets, spanning a total of five stimulus paradigms that embedded moral cues in increasing naturalistic settings: In the first dataset, the same subjects were exposed to controlled, decontextualized, visually presented moral vignettes and auditory movie summaries, as well as to more dynamic, audiovisual political attack advertisement. In dataset two, a different set of subjects freely viewed a theory-guided selection of audiovisual soap opera clips (Weber, 2008), featuring characters that violate or uphold moral norms and are subsequently punished or rewarded for their actions. In a third dataset, another set of participants freely watched an entire, full-length motion picture film. To evaluate the usefulness of naturalistic stimuli for studying moral cognition in a quantifiable fashion, we assessed the extent to which varying levels of stimulus naturalism lend themselves to a) identifying brain networks that undergird moral cognition; b) lead to shared or idiosyncratic neural responses in identified moral brain networks; and c) modulate the coherence between individual differences in morally relevant traits and behaviors and shared neural representations of moral stimuli.

Generally, we found that controlled moral vignettes are a reliable paradigm for identifying dissociable brain networks that undergird moral judgments. In turn, we demonstrated that moral brain circuits identified via controlled vignettes remain stable during the processing of more naturalistic moral narratives, providing evidence for both the ecological validity of controlled moral judgment tasks and the reliability of neural responses to moral transgressions in more naturalistic environments. Furthermore, we showed that the rich sampling space of naturalistic stimuli can advance the pluralistic study of moral cognition by illuminating how characters' identities (e.g., political affiliation), actions (e.g., violating or upholding moral norms), and consequences (e.g., rewards and punishments) dynamically interact with individual differences and drive intersubject synchrony in moral brain networks. At the same time, to prevent reverse inference errors (Poldrack, 2011) and to afford a more robust study of naturalistic moral perception, we encourage scholars to i) triangulate neural recordings during naturalistic perception with content-analytic and behavioral measurements that tap into morally relevant stimulus features, and ii) to take into account individuals differences in empathy, moral intuition profiles, and political ideology that interact with specific, morally relevant contextualized content features. Adhering to these practices likely boosts the advantages of naturalistic stimuli for interrogating how moral cognition dynamically evolves during narrative processing, is modulated by individual differences, and shapes character judgment, story engagement, and narrative appeal.

To elaborate on these findings, using classic brain mapping techniques, we demonstrated that processing moral vignettes elicits neural activation in a dissociable brain network including temporoparietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), Precuneus (PC), and superior temporal sulcus (STS). Given the common reference of these regions in the extant moral neuroscience literature Buckholtz & Marois, 2012; FeldmanHall & Mobbs, 2015; Greene & Haidt, 2002; Parkinson et al., 2011; Wasserman et al., 2017; Young & Dungan, 2012), we encourage future researchers to use moral vignettes as a standardized "localizer" for identifying regions that compose the moral brain. Furthermore, by applying intersubject synchrony analysis (Hasson et al., 2004; Nastase et al., 2019), we showed that moral violations embedded in auditory short stories, political attack advertisements, and soap opera clips elicited significant, stimulus-driven neural activation in the same ROIs as identified by the moral vignettes, providing converging evidence that moral transgressions, even when embedded in more ecologically-valid, but potentially noisy input streams, have reliable neural correlates.

Yet, precisely because narratives contain a much larger content feature space that not only consists of isolated moral actions as controlled tasks do, we stress that neural activation in moral brain networks during the processing of narratives may not necessarily index the *moral* judgment of characters and their actions. Thus, to lower the likelihood of reverse inference errors (Poldrack, 2011) when using naturalistic stimuli in moral neuroscience, we employed online (dynamic) and static (post-hoc) behavioral measurements that assessed subjects' moral judgments. Our behavioral results demonstrate that audiences indeed judged moral transgressions and their consequences in a predicted fashion, suggesting that the underlying, time-locked neural computations during narrative exposure indeed reflect moral cognitions.

At the same time, openly available, naturalistic fMRI datasets (Aliko et al., 2020; Nastase, Liu, Hillman, Zadbood, et al., 2020) frequently do not contain behavioral measurements that interrogate moral cognitions. As our secondary analysis of a full length movie matched this scenario, we triangulated neural activation patterns during movie viewing with validated computational content analysis of moral content features (Hopp et al., 2021) and continuous response measurements executed by a separate set of coders. Our results demonstrated that behavioral engagement ratings were positively correlated with both the story's underlying moral conflicts and the neural activation patterns in moral ROIs of fMRI participants. This finding points to the promising role of naturalistic moral narratives over short, decontextualized moral judgment tasks to advance media psychological theories that postulate linkages between a story's moral conflict pattern (Hopp et al., 2021), synchronization of audiences (Weber et al., 2013), and narrative appeal (Tamborini et al., 2013; Weber et al., 2008). In addition, these results may also encourage researchers who wish to use narrative stimuli for studying moral cognition to utilize moral content analysis pipelines and crowd-sourced ratings of narrative engagement as means to triangulate how recorded neural activations ebb and flow with a story's moral content and modulate narrative enjoyment.

Furthermore, speaking to the power of naturalistic narratives for synchronizing neural responses, we observed that intersubject neural similarity, particularly in TPJ and PC, increased continuously from moral vignettes to auditory movie summaries and political attack advertisements. As the same subjects were exposed to these stimuli, this is compelling evidence that more dynamic, multimodal, and engaging moral narratives are indeed more capable of synchronizing audiences' brains compared to static and decontextualized paradigms (Finn, 2021; Finn & Bandettini, 2021). Crucially, this variance in intersubject synchrony across different stimuli, but similar subjects, afforded us to examine whether intersubject representational similarities in moral brain networks come more into focus via a) highly controlled tasks that elicit minimal average intersubject synchrony and leave more room for specific subject pairs to be more or less correlated with one another according to variations in traits and behavior; or b) moral stimuli that strongly synchronize subjects' neural activity and boost signal-to-noise for individual differences, since even though these stimuli may reduce overall intersubject variability, the residual variation may be more stable and trait-like (Finn et al., 2017; Vanderwal et al., 2017). Here, our findings largely support Finn and colleagues' (2020) assertion that "there may not be a single 'best' stimulus for studying individual differences [in moral cognition]; rather, the most appropriate stimulus may depend on the specific [moral traits and] behavior(s) of interest." (p. 20). Accordingly, we did not find that intersubject RSA values increased equally for all traits and behaviors across increasingly naturalistic stimuli. In contrast, we found that individual differences in neural responses become more pronounced based on the specific content and context-based cues of the stimulus. For moral vignettes, we find that variation in general sensitivity to distressing actions resulted in the highest intersubject RSA values. Hence, highly controlled, decontextualized experimental stimuli may generally be better suited to study individual differences in domain-general personality traits (e.g., the big five, see Finn et al., 2020) rather than morally relevant traits that interact with contextualized content features, such as the identities and intentions of characters. In line with this argument, processing dynamic auditory movie summaries where the main character violates a moral foundation increased the coherence between individuals' overall similarity in moral intuition salience and intersubject neural activations. This renders contextualized narratives superior to controlled moral vignettes for studying how individual differences that particularly pertain to moral sensitivities drive neural representations. As expected, for political attack advertisements, individuals' similarity in political orientation yielded the highest intersubject RSA values in contrast to other stimuli, traits, and behaviors. Hence, dynamic political attack advertisements in which candidates commit contextualized moral transgressions may be a promising future stimulus for researchers interested in examining how neural polarization processes operate in the moral brain and are shaped by individuals' political ideology (Baar et al., 2021; Leong et al., 2020). Finally, to understand how the moral mind parses contextualized consequences of moral and immoral actions, our analysis of audiovisual soap opera clips highlighted that intersubject variability in continuous response measurements yielded the highest coherence with neural synchronies in scenarios where moral or immoral characters are punished. Hence, compared to more controlled, decontextualized paradigms, our analysis of naturalistic soap operas signals that responses to observed punishment are primarily driven by a desire to both provide justice for those who have been wronged and second to reprimand those who violate expected moral norms.

Limitations and Future Directions

Although this work has highlighted the opportunities of both controlled and naturalistic moral narratives for advancing our understanding of the moral brain, our analyses have limitations that warrant future research. One limitation of the present study is the rather coarse-grained spatial scale – 50 parcels – with which neural activation and intersubject representational similarities were linked across stimuli. While selecting the optimal spatial scale for multivariate feature selection in fMRI remains an open discussion (Jolly & Chang, 2021), mounting evidence points to the fact that idiosyncratic variations in traits and behavior are encoded at a much finer spatial resolution (Feilong et al., 2018). Yet, increasing spatial granularity by using a higher-resolution brain parcellation or even relying on a searchlight or voxelwise approach (Wasserman et al., 2017) significantly increases the computational costs of conducted analyses and necessitates stricter control of the familywise error rate. Nevertheless, future studies should connect to the herein described analyses by means of a more fine-grained measurement of neural activation patterns. Here, utilizing voxelwise or sphere-based, rather than ROI-based activation maps generated by controlled moral vignettes may serve as a healthy middle point that maintains informational specificity at a manageable computational scale. A related point of discussion is concerned with further improving the signal generated by individual variation while factoring out group variance. For instance, shared-response modeling and hyperalignment have been shown to increase sensitivity to individual differences (Feilong et al., 2018). Future work on individual differences in moral cognition should investigate whether these approaches increase sensitivity to brain-behavior relationships in the IS-RSA framework. Here, creating shared response spaces for the same subject, but across multiple datasets (Nastase, Liu, Hillman, Norman, et al., 2020) may be a particularly promising endeavor to boost individual signal.

Moreover, we herein largely treated the moral content features embedded in naturalistic stimuli as block-designs, where we concatenated and grouped together stimuli that featured common moral scenarios. Future work may triangulate this approach with more fine-grained event-related designs using content analysis to identify time segments where the moral action of interest occurred, who the actor was, if the action was intentional, and what the subsequent consequences of the action were. Using forward inference, these event codes could be used in classic encoding models to examine how stable or variable the neural signatures of moral transgressions are across naturalistic contexts. Especially when coupled with certain covariates of stimulus features (e.g., luminance, visual flow, acoustic properties (Lahnakoski et al., 2012), neural encodings of morality could be more robustly crystalized. Second, as demonstrated per the movie viewing paradigm, combining theoretically-informed content-analysis with dynamic intersubject synchrony approaches (Simony et al., 2016; Weber, 2008) may reveal whether the brain-trait signal indeed increases (or fades) during morally-salient narrative moments.

Lastly, this study primarily related intersubject variability in neural representations to moral judgments recorded immediately after stimulus exposure in a laboratory setting. Future research could thus extend this study by testing whether variability in neural responses to either controlled or naturalistic stimuli is a better predictor of subsequent real-world behavior(s), including similarity in moral sentiment expression on social networks, voting preferences, or movie ratings.

Conclusion

Taken together, this work has highlighted the challenges and opportunities of using highly controlled, decontextualized moral judgment tasks as well as more naturalistic, contextualized stimuli for illuminating the neural basis of moral cognition. In particular, we emphasized the usefulness of controlled moral vignettes for studying the dissociable brain networks that undergird moral judgments. Moreover, we demonstrated that moral brain circuits identified via controlled vignettes remain stable during the processing of more naturalistic moral narratives, providing evidence for the ecological-validity of experimentally controlled moral judgment tasks and the reliability of more naturalistic, contextualized narratives to elicit neural responses in moral brain networks. Furthermore, we showed that the rich sampling space of naturalistic stimuli can advance the pluralistic study of moral cognition by illuminating how characters' identities (e.g., political affiliation), actions (e.g., violating or upholding moral norms), and consequences (e.g., rewards and punishments) dynamically interact with individual differences and drive intersubject synchrony in moral brain networks. Finally, if researchers enrich neural recordings obtained during naturalistic moral perception with theory-driven content-analytic and behavioral ratings, narrative stimuli are a boon for studying how moral cognition evolves during narrative processing, is modulated by individual differences, and shapes character judgment, story engagement, and narrative appeal.

Methods

Description of Datasets

Moral Vignettes, Narrated Movie Summaries, and Political Attack Advertisements

The first three datasets (moral vignettes, narrated movie summaries, and political attack advertisements) are drawn from a previously conducted study that includes 64 participants recruited from communication courses at the University of California, Santa Barbara. Before undergoing MRI, participants completed a battery of questionnaires to assess their basic socio-moral (Moral Foundations Questionnaire, (Graham et al., 2011); Society Works Best, (Smith et al., 2011); Empathy, (Davis, 1980) and political attitudes (e.g. political behavior, knowledge and affiliation). Participants then underwent a first fMRI scan while completing the Moral Foundation Vignettes (MFV; Clifford et al., 2015). The MFV span 120, one-sentence descriptions (14-17 words) detailing the violation of one (and only one) of eight categories: Emotional care, physical care, fairness, loyalty, liberty, authority, purity, and social norms (control), with 15 vignettes in each category (Table 4).

Table 4. *Examples of Vignettes for the Eight Types of Violations.*

Moral Foundation	Example Vignette
Care Physical	You see a woman swerving her car in order to intentionally run over a squirrel.
Care Emotional	You see a girl saying that another girl is too ugly to be a varsity cheerleader.
Fairness	You see a soccer player pretending to be seriously fouled by an opposing player.
Liberty	You see a public leader on TV trying to ban the wearing of hooded sweatshirts.
Authority	You see a boy spray-painting anarchy symbols on the side of the police station.
Loyalty	You see a man leaving his family business to go work for their main competitor.
Purity	You see a man having sex with a frozen chicken before cooking it for dinner.
Social Norm (Control)	You see a woman eating dessert before her main entrée arrives on the table.

These vignettes were organized in an event-related design, pseudorandomly distributed over three approximately 8-minute functional runs (5 in each of the 3 runs). Subjects viewed one vignette at a time and were instructed to vividly imagine the described scene. While the vignette was on the screen, they were asked to make a judgment of how morally wrong the action described in the vignette was (1 = *not at all morally wrong* to 4 = *extremely morally wrong*). After 8 seconds, the vignette disappeared but the scale remained on screen and participants could respond during the inter-trial interval (ITI). ITIs were on average 4 seconds long, with a jitter of +/- 2.16 seconds (jitter length was calculated so that each trial would begin at exactly the beginning of the scanner's collection of the next volume).

Thereafter, participants listened to ten professionally narrated and previously validated (Tamborini et al., 2013) movie plots summaries (mean duration = 45sec) while undergoing fMRI. In each plot summary (2 per moral foundation), the main character (a) violates one of the five moral foundations while adhering to all other foundations during the "background" part of the story and (b) is either punished or rewarded for their moral transgressions during the “outcome” part of the story. Trial order was randomized and counterbalanced between subjects so that the same order of plots was given to a pair of subjects, but with the alternative outcomes. After listening to a summary, subjects rated, on a scale of 1–4, the degree to which they would like to see the movie. Subsequent to undergoing fMRI, subjects listened to each plot summary again (randomized order) and rated the degree to which they perceived (a) the main character’s actions to be moral–immoral, (b) the outcomes that befell the character to be good–bad, (b) the behavior of the character to uphold–violate each of five moral foundations.

In a third fMRI session, subjects watched political campaign attack ads (11 ads attacking Trump, 11 ads attacking Clinton). The length of the average ad was 38.35 seconds (range 29.26–78.20). Using an MRI-safe button box, subjects were instructed to press “2” whenever the perceived ad made a statement they liked and “1” whenever it made a statement they did not like. After watching each ad participants were also asked to rate the ad on a scale from 1–4 on how convincing the ad was. Time between each ad was 5 seconds, with +/–2 seconds of jitter during which the response was solicited. The total duration of the entire experimental run was ~16.5 minutes. The order of attack ads was generated using repeated history with a lookback of 1. For example, an anti-Trump ad appeared the same amount of times before another anti-Trump ad as an anti-Clinton ad. Then the order of ads was randomized within this constraint: a reverse order was presented to the next subject of same gender and political affiliation. Finally, each sequence was presented once to a member of each of the 6 main participant groups (Male/Female)x(Republican/Democrat /Unaffiliated). After the fMRI session, subjects re-watched all political ads and after each ad answered a battery of questions probing their reaction to the ad, including how strongly the ad was making them want to vote for a candidate, emotional impact, whether a moral violation was perceived in the different domains and whether the subject had seen the ad before. The sequence of ads in the survey was random and different from their sequence during fMRI.

All previously described MRI data was collected on a Siemens Magnetom Prisma 3T MRI following Human Connectome Project guidelines (Ugurbil et al., 2013). The blood-oxygen-level dependent (BOLD) contrast was measured using a multiband echo planar gradient sequence (TR = 720 ms, TE = 37 ms, FA = 52 degrees, FOV = 208 mm, acceleration factor = 8). Volumes consisted of 72 interleaved slices (2 mm isotropic) acquired with an angle of ~20^{o} relative to the AC-PC plane, so that the slices are acquired more dorsally near the eyes relative to the back of the brain (in that fashion we were able to acquire the entire brain volume including the cerebellum for every subject). High-resolution T1-weighted whole brain acquisitions were collected prior to functional image acquisition (TR = 2500 ms, TE 2.22 ms, FA = 7 degrees, FOV = 241 mm, .9 mm isotropic resolution).

Soap Opera Clips

Data for the soap opera clips was collected from a previously conducted study spanning 28 healthy female participants (Weber, 2008). Upon entering the lab, participants provided information on demography, handedness, medical history, trait empathy (Davis, 1980) and personality as indexed by (Eysenck & Eysenck, 1993) and the Ten Item Personality Index (TIPI, Gosling et al., 2003). Participants then watched 15 pre-selected scenes from the show *Days of Our Lives*–a popular daytime US network television soap opera–while undergoing fMRI. Notably, these scenes have been analyzed for the presence of moral arcs (who, what, whom, when, why/what intention, and consequence) using a theoretically-driven, methodical content analysis (Weber, 2008). Furthermore, the selection of scenes was based on a survey among an independent sample of 547 female undergraduate students (average was 20.1 years; SD = 1.09). The sample resembled both the main target group of the television show and the sample for the present study. Evaluations included the moral valence of character's behaviors (i.e., did the character act morally or immorally) and the valence of story outcomes for characters (i.e., whether moral or immoral behaviors were punished or rewarded). Scene selection was based on the presence of a clearly identifiable main character with either high, neutral, or low ratings on morality and clearly identified positive, neutral, or negative outcomes. Due to the nature of the programming, most moral behaviors violated or upheld the fairness/reciprocity norm. That is, immoral characters attempted to socially sabotage other people or attempted to cheat them in some manner, whereas moral characters sacrificed time and effort to help others. Outcomes were also mainly social in nature, such that in the positive outcome scenes, characters were socially rewarded by receiving good news or succeeding in their social machinations, whereas in the negative outcome scenes, characters were socially punished through being rejected or suffering emotional harm. Based on these categorizations, the 15 selected scenes were grouped into the following five experimental groups (3 per group): (1) perception of main character's behaviors as moral & perception of story outcomes for the main character as positive; (2) immoral behavior & negative outcome; (3) moral behavior & negative outcome; (4) immoral behavior & positive outcome; (5) neutral behavior (perception of behavior as neither moral nor immoral) & neutral outcome (perception of story outcome as neither negative nor positive). The presentation order of the scenes was fully randomized across participants. Scenes were presented on a 640x480 color LCD in the magnet (as hood mount) and synchronized with the MR-signal. Participants wore MRI headphones in the scanner to hear the specific video clip. The entire scanning procedure was divided in 3 runs of 18 minutes each, resulting in a total scan time of 54 minutes. Between clips, participants were shown a blue screen with a centered cross for 30 seconds which served as a non-active baseline.

Cortical activity was measured as the BOLD effect. BOLD contrast was obtained with a gradient-echo EPI sequence (General Electric scanner; field strength of 3 Tesla; whole brain coverage with 30 interleaved slices; slice size 4mm with 0.4mm gap; TR = 2000ms; TE = 27.2 ms; flip angle = 77°, field of view 22 × 22 cm2, matrix size 64 × 64;). For reference, we acquired anatomical brain images of each participant between the first and the second run of the fMRI procedure. Raw DICOMs were organized according to the *Brain Imaging Data Structure* (BIDS; Gorgolewski et al., 2016) and then minimally pre-processed using *fMRIprep* (Esteban et al., 2019).

After completing fMRI, participants were seated in front of a PC and watched all 15 scenes again and evaluated their narrative enjoyment of each clip using continuous response measurement. Participants' moment-to-moment enjoyment of each scene was assessed via continuous response measurement (CRM; Biocca et al., 1994). Using the Media Monitor software, participants were instructed to re-watch each of the 15 scenes while continually moving a slider to rate how much they enjoyed each moment on screen (*I do not enjoy it at all – I enjoy it very much)*. Due to software malfunctions, CRM data from four participants had to be discarded. Lastly, for each viewed clip, participants additionally rated main characters’ moral behaviors and outcome valence.

Full-Length Movie

We use data from the recently released *naturalistic neuroimaging database* (NNDb v.1.0; (Aliko et al., 2020). The NNDb spans 84 participants watching 10 full-length movies from 10 genres. Notably, we only focused on the movie *500 Days of Summer* as this movie was viewed by a total of 20 subjects, whereas other movies were only viewed by 8 subjects at a time. In addition, we obtained the movie's publicly available, manually-corrected, high-quality movie script in extended markup language (XML) format (Gorinski & Lapata, 2015) as well as its subtitles from https://www.opensubtitles.org.

fMRI Preprocessing

For each described dataset, raw DICOMs were organized according to the *Brain Imaging Data Structure* (BIDS; Gorgolewski et al., 2016) and then minimally pre-processed using *fMRIprep* (Esteban et al., 2019)–a robust tool to prepare task-based fMRI data for statistical analysis and circumventing the reproducibility concerns of fMRI preprocessing steps (Esteban et al., 2020). After preprocessing with fmriprep, we postprocessed the data according to established guidelines for naturalistic imaging data (https://naturalistic-data.org/content/Preprocessing.html).

Specifically, we smoothed the data (fwhm=6mm) and performed basic voxelwise denoising using a General Linear Model. This entailed including the 6 realignment parameters, their squares, their derivatives, and squared derivatives. We also included dummy codes for spikes identified from global signal outliers and outliers identified from frame differencing (i.e., temporal derivative). We chose to not perform high-pass filtering and instead include linear & quadratic trends, and average CSF activity to remove additional physiological and scanner artifacts. Note that the movie from the NNdb had already been pre- and postprocessed and organized into BIDS by the authors, so no further processing was applied.

Parcellation

All imaging analyses were conducted over a set of 50 parcels (parcellation available at http://neurovault.org/images/39711). The parcellation was created by performing a whole-brain parcellation of the coactivation patterns of activations across over 10,000 published studies available in the Neurosynth database (Yarkoni et al., 2011). The use of a parcellation scheme has several advantages over the more conventional voxelwise and searchlight approaches. First, it is several orders of magnitude less computationally expensive as analyses are performed only 50 times compared to 352,000 voxels. Second, the parcels are nonoverlapping and contain bilateral regions that reflect functional neuroanatomy, whereas a searchlight approach is limited to local spheres that do not adapt to different areas of cortex.

Analyses

All analyses were performed in a dedicated Anaconda environment using Python 3.7.9 (see https://github.com/medianeuroscience/diss_fhopp/blob/main/conda_env.yml). Unless otherwise noted, all fMRI analyses were performed using the *NLTools* package version 0.4.4 (https://github.com/cosanlab/nltools).

Vignettes General Linear Model

The moral vignettes were used to determine ROIs that are preferentially activated during the judgment of moral violations. The functional data was analyzed using general multilevel linear modeling procedures (GLM; Beckmann et al., 2003). First level GLMs for each subject included explanatory variable time courses (EVs) for the control condition (social norm violations) and for each moral domain (gamma convolution = 6 s, SD = 3) and temporal derivatives of each EV. For each trial a 5-second window (starting 5 seconds from when the vignette appears on the screen) of volumes of interest was included as an event. Planned contrasts then modeled activation unique to each moral domain (e.g., fairness vs.

control). These contrasts were then analyzed at the second level using a mixed effects analysis. Resulting statistical parametric maps were thresholded using FDR ($q < .05$). Parcels that survived FDR were then included in all subsequent analyses.

IS-RSA Vignettes

We first computed each participant's moral vs. social vignette contrast maps (one per moral domain, 7 total) by subtracting each mean (averaged across three runs) moral GLM beta map from the mean social GLM beta map. We then divided these moral–social subject-level beta maps into 50 ROIs using the described parcellation. Next, we created a pairwise correlation matrix for each ROI (the "ROI similarity matrices"). Correlation is a useful metric that can accommodate data that is on different scales, which is important when comparing different participants' beta maps. We also created similarity matrices for each trait and behavior measure. For empathy, moral intuition salience, and the vignette wrongness ratings, we computed the mean of every subject pair's rank on a given scale, normalized by the highest possible rank. Notably, for moral intuition salience and wrongness ratings, we did this separately for each foundation to test whether individual differences in moral traits and behavior predict neural similarity in a foundation-specific fashion. For political orientation, we computed the pairwise Euclidean distance and then converted it to similarity. Accordingly, these "individual difference similarity matrices" captured the similarity between participants' in their traits and behavior.

For each moral–social vignette contrast, we then computed the correlation between each ROI similarity matrix and the individual difference similarity matrices using Spearman's rank-order correlations on the lower triangle of the matrices. To obtain significance levels of the resulting Spearman's rhos, we used a Mantel permutation test (Mantel, 1967; Nummenmaa et al., 2012), in which both the rows and columns of one subject by subject similarity matrix are shuffled and the Spearman's rhos between both correlation matrices is recomputed. This procedure was repeated 10,000 times to generate a null distribution of rank correlations which was used to compute *p*-values based on a one-tailed test with the alternative hypothesis of correlations greater than 0 for each ROI (Nili et al., 2014). These Monte Carlo *p*-values were Bonferroni-corrected by multiplying them by the number of parcels (50). All *p*-values that remained below 0.05 after this correction were taken to indicate a significant association between individual difference similarity and ROI representation similarity.

ISC and IS-RSA for Movie Plot Summaries

We used intersubject correlation (ISC) to examine the reliability of neural dynamics in response to the movie summaries across individuals (Cohen et al., 2017; Hasson et al., 2004; Nastase et al., 2019). For each participant, we extracted the timepoints when participants were listening to the summaries of a particular foundation (two per foundation) and separately concatenated all of the summaries into five separate conditions (one per moral foundation). We then separately extracted the mean time course for each condition with 50 ROIs – using the same parcellation scheme as before. For each summary type, we separately computed the pairwise correlation between participants' mean time-course in each ROI producing a subject by subject correlation matrix for each ROI. To compare the spatial patterns of mean ISC between foundations, we computed the mean of the lower triangle of the subject by subject correlation matrix for each ROI and then correlated these mean values between each foundation. This correlation provides an estimate of the similarity of individual variation across brain regions between each summary.

We performed IS-RSA to identify brain regions with temporal dynamics that exhibited similar patterns of intersubject variability to either trait-level (i.e., political orientation, empathy, moral intuition salience) or behavioral-level (i.e., foundation and character ratings) individual differences. The overall analytical framework was identical to the IS-RSA previously reported for the vignettes, with the exception that the condition-specific ROI similarity matrices were obtained through the ISC analysis above.

ISC and IS-RSA for Political Attack Advertisements

We again performed ISC analysis to examine the reliability of neural dynamics in response to the political attack advertisements across individuals. For each participant, we extracted the timepoints when participants were watching either anti-Trump or anti-Clinton advertisements and separately concatenated all of the clips into two conditions (anti-Trump or anti-Clinton). We then separately extracted the mean time course for each condition within the same 50 ROIs as before. For each condition type, we separately computed the pairwise correlation between participants' mean time-course in each ROI producing a subject by subject correlation matrix for each ROI. To compare the spatial patterns of mean ISC between conditions, we computed the mean of the lower triangle of the subject by subject correlation matrix for each ROI and then correlated these mean values between each condition. This correlation provides an estimate of the similarity of individual variation across brain regions between each condition.

The IS-RSA for the political attack advertisements followed the identical steps as the IS-RSA reported for the movie summaries. We again searched for regions of the brain that showed similar variability as measured via ISC to either trait-level (i.e., political orientation, empathy, moral intuition salience) or behavioral-level (i.e., perceived moral foundation and morality ratings) individual differences.

ISC and IS-RSA for Soap Opera Clips

We used intersubject correlation (ISC) to examine the reliability of neural dynamics in response to the soap opera clips across individuals (Cohen et al., 2017; Hasson et al., 2004; Nastase et al., 2019). For each participant, we extracted the timepoints when participants were freely viewing clips pertaining to each of the five conditions (three clips per condition) and separately concatenated all of the clips into five separate conditions. We then separately extracted the mean time course for each condition with our 50 ROI parcellation. For each condition, we separately computed the pairwise correlation between participants' mean time-course in each ROI producing a subject by subject correlation matrix for each ROI. To compare the spatial patterns of mean ISC between conditions, we computed the mean of the lower triangle of the subject by subject correlation matrix for each ROI and then correlated these mean values between each condition. This correlation provides an estimate of the similarity of individual variation across brain regions between each condition.

We performed IS-RSA to identify brain regions with temporal dynamics that exhibited similar patterns of intersubject variability to either trait-level or behavioral-level individual differences. The overall analytical framework was identical to the IS-RSA reported for the previous datasets. On the trait level, we computed participants' overall similarity in empathy via an NN-itemwise model and again used an AnnaK model for each empathy subdimension. Likewise, we computed two NN-itemwise models that captured similarity in personality as measured by Eysenck and the TIPI. On the behavioral level, we computed intersubject similarity in moral judgment for character morality and outcome valence using an AnnaK model on the average ratings per condition (three ratings per condition) and an NN-itemwise model capturing the correlation of ratings across clips per condition. Lastly, for each condition, we created a similarity matrix denoting the pairwise, intersubject correlation for continuous response measurements (CRMs) across the three condition-specific, concatenated clips.

Movie Annotation Pipeline

We first obtained the screenplay for the movie *500 Days of Summer* from *Scriptbase J* (Gorinski & Lapata, 2015). Properly formatted movie scripts allow the distinction between several structural screenplay features such as scenes and stage directions, action descriptions, characters, and dialogue. Hence, we computationally parsed the script using a pipeline developed in (Hopp et al., 2020) to extract the labels for every scene onset. Next, to align the scenes of a movie script with the corresponding motion picture, two research assistants jointly viewed the movie while labeling the timestamps corresponding to the onsets of each extracted screenplay scene. Additional scenes that were not present in the screenplay were also recorded, while scenes that were missing in the movie were discarded and scenes spanning multiple scenes in the screenplay grouped together. Thereafter, we assessed the moral signal of a scene by computationally detecting morally relevant words in the subtitles of the movie. To this end, we first obtained the subtitles for the movie from https://www.opensubtitles.org/ and then again computationally parsed the downloaded file. As a result, every line of this file contains a unique subtitle element, with the corresponding text as well as beginning and ending of the text on screen. Next, we aligned the subtitle time codes with the scene onset codes to learn what subtitles were present during a given scene.

Thereafter, we extracted the moral words in each subtitle file by using the extended Moral Foundations Dictionary (eMFD; (Hopp, Fisher, Cornell, et al., 2020) The eMFD is built on a crowd-sourced content analysis and contains a total of 3,270 words. Each word in the eMFD is assigned five probabilities that denote the association with each of the five moral foundations. Using the eMFDscore software[2] and the *bow single vice-virtue* scoring option, for each scene, we obtained ten probabilities that reflect the degree to which any one moral foundation is present and upheld or violated in a scene.

[2] (https://github.com/medianeuroscience/emfdscore)

References

Adebimpe, A., Bassett, D. S., Jamieson, P. E., & Romer, D. (2019). Intersubject Synchronization of Late Adolescent Brain Responses to Violent Movies: A Virtue-Ethics Approach. *Frontiers in Behavioral Neuroscience*, *13*. https://doi.org/10.3389/fnbeh.2019.00260

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, *7*(1), 347. https://doi.org/10/gjvjnq

Amit, E., & Greene, J. D. (2012). You See, the Ends Don't Justify the Means: Visual Imagery and Moral Judgment. *Psychological Science*, *23*(8), 861–868. https://doi.org/10/f4pwvd

Baar, J. M. van, Halpern, D. J., & FeldmanHall, O. (2021). Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences*, *118*(20). https://doi.org/10.1073/pnas.2022491118

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of Real-World Event Schemas during Narrative Perception. *The Journal of Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Bandura, A. (2001). Social Cognitive Theory of Mass Communication. *Media Psychology*, *3*(3), 265–299. https://doi.org/10/cpb6nq

Berman, J. Z., & Kupor, D. (2020). Moral Choice When Harming Is Unavoidable. *Psychological Science*, 095679762094882. https://doi.org/10.1177/0956797620948821

Biocca, F., David, P., & West, M. (1993). Continuous Response Measurement (CRM): A Computerized Tool for Research on the Cognitive Processing of Media Messages. *A. Lang (Ed.)*, 15–65.

Bowman, N. D., Jöckel, S., & Dogruel, L. (2012). A question of morality? The influence of moral salience and nationality on media preferences. *Communications: The European Journal of Communication Research*, *37*(4), 345–369.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661. https://doi.org/10.1038/nn.3087

Chang, C., Lazaridi, C., Yeshurun, Y., Norman, K. A., & Hasson, U. (2020). Relating the past with the present: Information integration and segregation during ongoing narrative processing. *BioRxiv*.

Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K. M., Greenstein, N., Chen, P.-H. A., & Manning, J. R. (2021). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science Advances*, *7*(17), eabf7129. https://doi.org/10/gjs4pw

Chen, P.-H. A., Jolly, E., Cheong, J. H., & Chang, L. J. (2019). Inter-subject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *BioRxiv*, 726570. https://doi.org/10.1101/726570

Chen, P.-H. A., Jolly, E., Cheong, J. H., & Chang, L. J. (2020). Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, *216*, 116851. https://doi.org/10/ggtvsx

Cheong, J. H., Xie, T., Byrne, S., & Chang, L. J. (2021). Py-Feat: Python Facial Expression Analysis Toolbox. *ArXiv:2104.03509 [Cs, Eess]*. http://arxiv.org/abs/2104.03509

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178–1198. https://doi.org/10.3758/s13428-014-0551-2

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin Modulates Behavioral Reactions to Unfairness. *Science*, *320*(5884), 1739–1739. https://doi.org/10.1126/science.1155577

Davis, M. H. (1980). *A multidimensional approach to individual differences in empathy*.

Dawson, K. J., Han, H., & Choi, Y. R. (2021). How are moral foundations associated with empathic traits and moral identity? *Current Psychology*. https://doi.org/10/gm8drq

Decety, J. (2021). Why Empathy Is Not a Reliable Source of Information in Moral Decision Making. *Current Directions in Psychological Science*, 09637214211031943. https://doi.org/10.1177/09637214211031943

Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General*, *147*(11), 1713–1727. https://doi.org/10/gfh32r

Eden, A., Daalmans, S., & Johnson, B. K. (2017). Morality Predicts Enjoyment But Not Appreciation of Morally Ambiguous Characters. *Media Psychology*, *20*(3), 349–373. https://doi.org/10.1080/15213269.2016.1182030

Eden, A., Grizzard, M., & Lewis, R. J. (2011). Disposition development in drama: The role of moral, immoral and ambiguously moral characters. *International Journal of Arts and Technology*, *4*(1), 33–47. https://doi.org/10.1504/IJART.2011.037768

Eden, A., Oliver, M. B., Tamborini, R., Limperos, A., & Woolley, J. (2015). Perceptions of Moral Violations and Personality Traits Among Heroes and Villains. *Mass Communication and Society*, *18*(2), 186–208. https://doi.org/10.1080/15205436.2014.923462

Eden, A., Tamborini, R., Aley, M., & Goble, H. (2021). Advances in Research on the Model of Intuitive Morality and Exemplars (MIME). In P. Vorderer & C. Klimmt (Eds.), *The Oxford Handbook of Entertainment Theory* (pp. 230–249). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190072216.013.13

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384. https://doi.org/10/dhbqch

Eriksson, K., Simpson, B., & Strimling, P. (2019). Political double standards in reliance on moral foundations. *Judgment and Decision Making*, *14*(4), 440–454.

Feilong, M., Nastase, S. A., Guntupalli, J. S., & Haxby, J. V. (2018). Reliable individual differences in fine-grained cortical functional architecture. *NeuroImage*, *183*, 375–386. https://doi.org/10/gfjmz8

FeldmanHall, O., & Mobbs, D. (2015). A Neural Network for Moral Decision Making. In *Brain Mapping* (pp. 205–210). Elsevier. https://doi.org/10.1016/B978-0-12-397025-1.00180-9

FeldmanHall, O., Mobbs, D., & Dalgleish, T. (2014). Deconstructing the brain’s moral network: Dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *9*(3), 297–306. https://doi.org/10.1093/scan/nss139

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in Cognitive Sciences*. https://doi.org/10/gm2nvc

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, *235*, 117963. https://doi.org/10/gjsk99

Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., & Constable, R. T. (2018). Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature Communications*, *9*(1), 2043. https://doi.org/10/gdpcpb

Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*, 116828. https://doi.org/10.1016/j.neuroimage.2020.116828

Finn, E. S., Scheinost, D., Finn, D. M., Shen, X., Papademetris, X., & Constable, R. T. (2017). Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage*, *160*, 140–151. https://doi.org/10/gcj67z

Gantman, A., Devraj-Kizuk, S., Mende-Siedlecki, P., Van Bavel, J. J., & Mathewson, K. E. (2020). The time course of moral perception: An ERP investigation of the moral pop-out effect. *Social Cognitive and Affective Neuroscience*, *15*(2), 235–246. https://doi.org/10.1093/scan/nsaa030

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*(1), 22–29. https://doi.org/10.1016/j.cognition.2014.02.007

Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. *Trends in Cognitive Sciences*, *19*(11), 631–633. https://doi.org/10.1016/j.tics.2015.08.004

Garten, J., Kennedy, B., Hoover, J., Sagae, K., & Dehghani, M. (2019). Incorporating Demographic Embeddings Into Language Understanding. *Cognitive Science*, *43*(1), e12701. https://doi.org/10.1111/cogs.12701

Gorinski, P. J., & Lapata, M. (2015). Movie script summarization as graph-based scene extraction. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2015.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10/cmz

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Greene, J., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Grizzard, M., Fitzgerald, K., Francemone, C. J., Ahn, C., Huang, J., Walton, J., McAllister, C., & Eden, A. (2019). Validating the extended character morality questionnaire. *Media Psychology*, 1–24.

Grizzard, M., Francemone, C. J., Fitzgerald, K., Huang, J., & Ahn, C. (2020). Interdependence of Narrative Characters: Implications for Media Theories. *Journal of Communication*. https://doi.org/10.1093/joc/jqaa005

Grizzard, M., Huang, J., Fitzgerald, K., Ahn, C., & Chu, H. (2018). Sensing Heroes and Villains: Character-Schema and the Disposition Formation Process. *Communication Research*, *45*(4), 479–501. https://doi.org/10.1177/0093650217699934

Hasson, U. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hester, N., & Gray, K. (2020). The Moral Psychology of Raceless, Genderless Strangers. *Perspectives on Psychological Science*, *15*(2), 216–230. https://doi.org/10.1177/1745691619885840

Hinyard, L. J., & Kreuter, M. W. (2007). Using Narrative Communication as a Tool for Health Behavior Change: A Conceptual, Theoretical, and Empirical Overview. *Health Education & Behavior*, *34*(5), 777–792. https://doi.org/10/dw2nc2

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2020). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01433-0

Hopp, F. R., Fisher, J. T., & Weber, R. (2020). A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts. *Media and Communication*, *8*(3), 164–179. https://doi.org/10.17645/mac.v8i3.3155

Hopp, F. R., & Weber, R. (2020). The state-of-the-art and the future of functional magnetic resonance imaging in communication research. In K. Floyd & R. Weber, *The handbook of communication science and biology* (pp. 279–291). Routledge.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, *38*(1), 52–62.

Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PloS One*, *3*(8), e2939.

Jolly, E., & Chang, L. J. (2021). Multivariate spatial feature selection in fMRI. *Social Cognitive and Affective Neuroscience*, *16*(8), 795–806. https://doi.org/10/gkfwhc

Jordan, J. J., & Kouchaki, M. (n.d.). Virtuous victims. *Science Advances*, *7*(42), eabg5902. https://doi.org/10/gm7hdr

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, *7*(4), 393–402. https://doi.org/10/ff6h7z

Kelly, C., & O'Connell, R. (2020). Can Neuroscience Change the Way We View Morality? *Neuron*, *108*(4), 604–607. https://doi.org/10.1016/j.neuron.2020.10.024

Kleemans, M., Eden, A., Daalmans, S., van Ommen, M., & Weijers, A. (2017). Explaining the role of character development in the evaluation of morally ambiguous characters in entertainment media. *Poetics*, *60*, 16–28.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 5648–5653. https://doi.org/10.1073/pnas.1207992110

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Krakowiak, K. M., & Tsay-Vogel, M. (2015). The Dual Role of Morally Ambiguous Characters: Examining the Effect of Morality Salience on Narrative Responses. *Human Communication Research*, *41*(3), 390–411. https://doi.org/10.1111/hcre.12050

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Lahnakoski, J. M., Salmi, J., Jääskeläinen, I. P., Lampinen, J., Glerean, E., Tikka, P., & Sams, M. (2012). Stimulus-Related Independent Component and Voxel-Wise Analysis of Human Brain Activity during Free Viewing of a Feature Film. *PLOS ONE*, *7*(4), e35215. https://doi.org/10/gm8m3g

Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences*, *117*(44), 27731–27739. https://doi.org/10.1073/pnas.2008530117

Mar, R. A., & Oatley, K. (2008). The Function of Fiction is the Abstraction and Simulation of Social Experience. *Perspectives on Psychological Science*, *3*(3), 173–192. https://doi.org/10.1111/j.1745-6924.2008.00073.x

McIntosh, A. R. (2004). Contexts and catalysts. *Neuroinformatics*, *2*(2), 175–181. https://doi.org/10/fnm833

McKee, R. (1997). *Story: Style, structure, substance, and the principles of screenwriting*. Harper Collins.

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. https://doi.org/10.1093/scan/nsz037

Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A., & Hasson, U. (2020). Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, *217*, 116865. https://doi.org/10/gg9n5b

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2020). *Narratives: FMRI data for evaluating models of naturalistic language comprehension* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.12.23.424091

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, *184*, 161–170. https://doi.org/10.1016/j.neuroimage.2018.09.010

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, *10*(4), e1003553. https://doi.org/10/gfwjrv

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences*, *109*(24), 9599–9604. https://doi.org/10.1073/pnas.1206095109

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is Morality Unified? Evidence that Distinct Neural Systems Underlie Moral Judgments of Harm, Dishonesty, and Disgust. *Journal of Cognitive Neuroscience*, *23*(10), 3162–3180. https://doi.org/10.1162/jocn_a_00017

Pegado, F., Hendriks, M. H. A., Amelynck, S., Daniels, N., Bulthé, J., Masson, H. L., Boets, B., & Beeck, H. O. de. (2018). Neural representations behind 'social norm' inferences in humans. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-31260-5

Poldrack, R. A. (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. *Neuron*, *72*(5), 692–697. https://doi.org/10/b9jxkf

Raney, A. A. (2002). Moral Judgment as a Predictor of Enjoyment of Crime Drama. *Media Psychology*, *4*(4), 305–322. https://doi.org/10.1207/S1532785XMEP0404_01

Raney, A. A. (2004). Expanding Disposition Theory: Reconsidering Character Liking, Moral Evaluations, and Enjoyment. *Communication Theory*, *14*(4), 348–369. https://doi.org/10.1111/j.1468-2885.2004.tb00319.x

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv:1908.10084 [Cs]*. http://arxiv.org/abs/1908.10084

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Schein, C. (2020). The Importance of Context in Moral Judgments. *Perspectives on Psychological Science*, *15*(2), 207–215. https://doi.org/10.1177/1745691620904083

Schultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, *194*(3), 465–475. https://doi.org/10.1007/s00221-009-1721-9

Shenhav, A., & Greene, J. D. (2010). Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, *67*(4), 667–677. https://doi.org/10.1016/j.neuron.2010.07.020

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension | Nature Communications. *Nature Communications*, *7*(1), 12141. https://doi.org/10.1038/ncomms12141

Sinnott-Armstrong, W. (2016). The disunity of morality. *Moral Brains: The Neuroscience of Morality*, 331–354.

Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. *The Monist*, *95*(3), 355–377.

Smith, K. B., Oxley, D. R., Hibbing, M. V., Alford, J. R., & Hibbing, J. R. (2011). Linking Genetics and Political Attitudes: Reconceptualizing Political Ideology. *Political Psychology*, *32*(3), 369–397. https://doi.org/10/c7qsxb

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading Stories Activates Neural Representations of Visual and Motor Experiences. *Psychological Science*, *20*(8), 989–999. https://doi.org/10.1111/j.1467-9280.2009.02397.x

Tamborini, R. (2013). Model of intuitive morality and exemplars. In R. Tamborini (Ed.), *Media and the Moral Mind*. Routledge.

Tamborini, R., Eden, A., Bowman, N. D., Grizzard, M., Weber, R., & Lewis, R. J. (2013). Predicting media appeal from instinctive moral values. *Mass Communication and Society*, *16*(3), 325–346. https://doi.org/10.1080/15205436.2012.703285

Tamborini, R., & Weber, R. (2020). Advancing the model of intuitive morality and exemplars. In K. Floyd & R. Weber (Eds.), *The Routledge Handbook of Communication Science and Biology*. Routledge.

Tamir, D. I., Bricker, A. B., Dodell-Feder, D., & Mitchell, J. P. (2016). Reading fiction and reading minds: The role of simulation in the default network. *Social Cognitive and Affective Neuroscience*, *11*(2), 215–224. https://doi.org/10/f8h9cp

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint models of neural and behavioral data*. Springer.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1), 1483. https://doi.org/10.1038/s41467-019-09161-6

van Leeuwen, F., Koenig, B. L., Graham, J., & Park, J. H. (2014). Moral concerns across the United States: Associations with life-history variables, pathogen prevalence, urbanization, cognitive ability, and social class. *Evolution and Human Behavior, 35*(6), 464–471. https://doi.org/10.1016/j.evolhumbehav.2014.06.005

Vanderwal, T., Eilbott, J., Finn, E. S., Craddock, R. C., Turnbull, A., & Castellanos, F. X. (2017). Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage*, *157*, 521–530. https://doi.org/10.1016/j.neuroimage.2017.06.027

Vega, A. de la, Chang, L. J., Banich, M. T., Wager, T. D., & Yarkoni, T. (2016). Large-Scale Meta-Analysis of Human Medial Frontal Cortex Reveals Tripartite Functional Organization. *Journal of Neuroscience*, *36*(24), 6553–6562. https://doi.org/10.1523/JNEUROSCI.4402-15.2016

Voelkel, J. G., & Brandt, M. J. (2019). The Effect of Ideological Identification on the Endorsement of Moral Values Depends on the Target Group. *Personality and Social Psychology Bulletin*, *45*(6), 851–863. https://doi.org/10/gh635g

Wasserman, E. A., Chakroff, A., Saxe, R., & Young, L. (2017). Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis. *NeuroImage*, *159*, 371–387. https://doi.org/10.1016/j.neuroimage.2017.07.043

Weber, R. (2008). *Connectivity of brain regions during social interaction: Theory-based, event-related content analysis of continuous, semi-natural stimuli as paradigm in functional mangetic resonance imaging* [PhD Thesis].

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *12*(2–3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Weber, R., Popova, L., & Mangus, J. M. (2013). Universal morality, mediated narratives, and neural synchrony. *Media and the Moral Mind*, 26–42.

Weber, R., Tamborini, R., Lee, H. E., & Stipp, H. (2008). Soap Opera Exposure and Enjoyment: A Longitudinal Test of Disposition Theory. *Media Psychology*, *11*(4), 462–487. https://doi.org/10.1080/15213260802509993

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for Neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003

Wright, P., He, G., Shapira, N. A., Goodman, W. K., & Liu, Y. (2004). Disgust and the insula: FMRI responses to pictures of mutilation and contamination. *NeuroReport*, *15*(15), 2347–2351.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10/btqbtq

Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience*, *7*(1), 1–10. https://doi.org/10.1080/17470919.2011.569146

Young, L., & Saxe, R. (2008). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

Yudkin, D. A., Gantman, A. P., Hofmann, W., & Quoidbach, J. (2021). Binding moral values gain importance in the presence of close others. *Nature Communications*, *12*(1), 1–12. https://doi.org/10/gjz347

Zillmann, D., & Cantor, J. R. (1977). Affective responses to the emotions of a protagonist. *Journal of Experimental Social Psychology*, *13*(2), 155–165. https://doi.org/10/bt6hjw

Zillmann, D., & Vorderer, P. (2000). *Media entertainment: The psychology of its appeal*. Taylor & Francis.

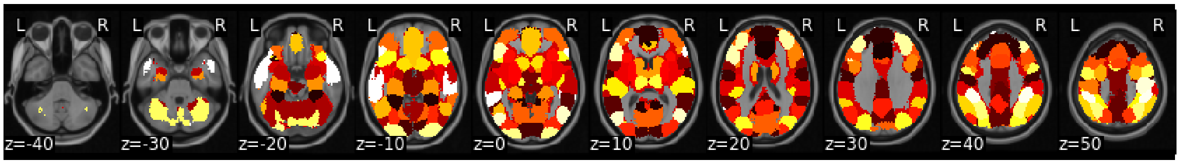
Zillmann, D. (2000). Basal morality in drama appreciation. *Moving Images, Culture, and the Mind*, 53–63.

Zillmann, D. (2002). Exemplification theory of media influence. In *Media effects* (pp. 29–52). Routledge.

Zillmann, D., & Cantor, J. R. (1976). *A disposition theory of humour and mirth.*

Supplemental Information

Neurosynth Parcellation (k = 50)



Controlled, Decontextualized Moral Foundation Vignettes

Multivoxel Pattern Classification of Moral Foundation Vignettes

We supplemented the encoding analyses of the moral foundation vignettes with a decoding approach using multi-voxel pattern analysis (MVPA). Specifically, we tested whether ROIs identified via our GLM represent vignettes of each condition differently. To this end, we first performed an initial temporal data reduction using univariate GLMs to create an average beta map of each participant's brain response to each of the eight vignette categories. We then used a leave-one-subject-out (LOSO) cross-validation procedure to evaluate the performance of our multivariate SVM model in classifying maps associated with each participant's response to a moral vignette category using data from the remaining 63 participants.

The trained SVM was able to discriminate (classification accuracy > 50%) between moral versus social norm violations in all ROIs identified by our localizer, with particularly high accuracy in posterior cingulate cortex (PCC), anterior vlPFC, and dmPFC, emphasizing the role of these ROIs in classifying an action as *morally* relevant (Figure 1). Classification accuracy across moral foundations was highest in PCC (Precuneus/superior LOC), suggesting that this region may play a superordinate role in discriminating different types of moral violations–a result consistent with extant work linking PCC activity to reasoning through different types of moral judgments (Greene et al., 2001) and hypothetical moral choices (FeldmanHall, Dalgleish, et al., 2012).

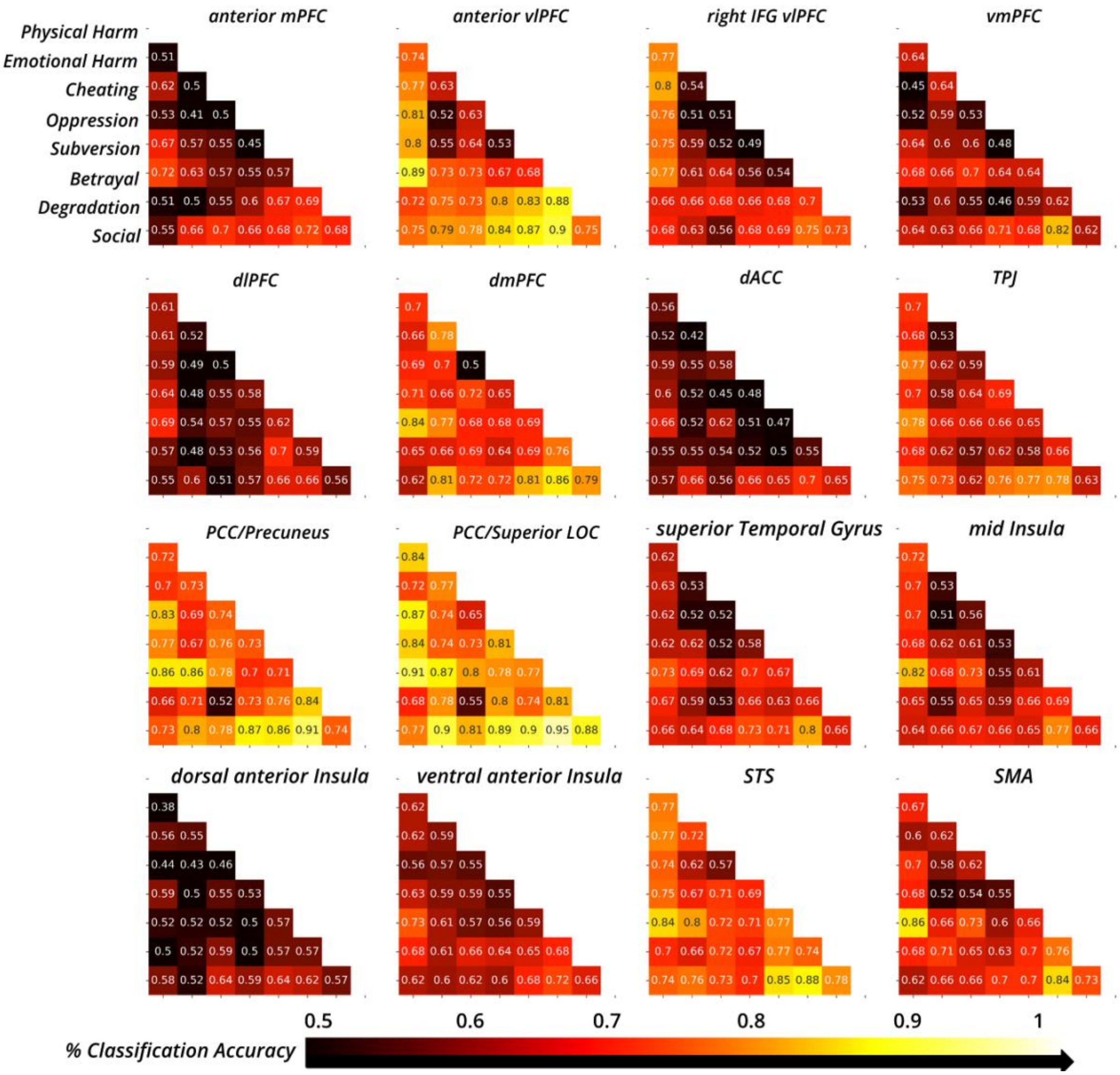


Figure 1. Between subject multivoxel pattern classification across moral vignette categories.

Representational Similarity Analysis

To gain more information about the representational geometry (Kriegeskorte & Kievit, 2013) of moral and social norm violations, we used representational similarity analysis (RSA; Kriegeskorte et al., 2008). We first obtained an activity estimate for each voxel and vignette item using massively univariate linear modeling (Kriegeskorte et al., 2008), creating a canonical beta map (item vs. average) for each of the 98 vignette items across our participants. For each ROI, the item-related activity patterns form the basis for computing the neural representational similarity matrices (RSM). We then created several candidate models to explain the representational pattern in our identified ROIs (Figure 2a, top).

Our conceptual, theoretically-derived models predicted greater similarity within versus between vignettes of the same condition; within moral versus social norm transgressions; within individualizing versus binding versus social norms; and individualizing versus binding as subcategories of moral versus social. In addition, we created two behavioral models capturing similarity in average moral wrongness ratings and response times across all vignettes. Using the sentence BERT (sBERT) transformer (Reimers & Gurevych, 2019), we also created a matrix denoting the semantic similarity of the vignettes in a high-dimensional sentence embedding space.

To explore the dimensions of the stimulus space that are most strongly reflected in the neural response patterns, we computed the group-average neural RSM for several candidate ROIs identified by our localizer and also for V1 due to its popular application in representational models (Figure 2a, bottom). When visually examining the neural RSMs, we noticed a salient distinction between moral versus social norm transgressions across ROIs, further supporting the notion that the brain represents moral versus social norm transgressions differently.

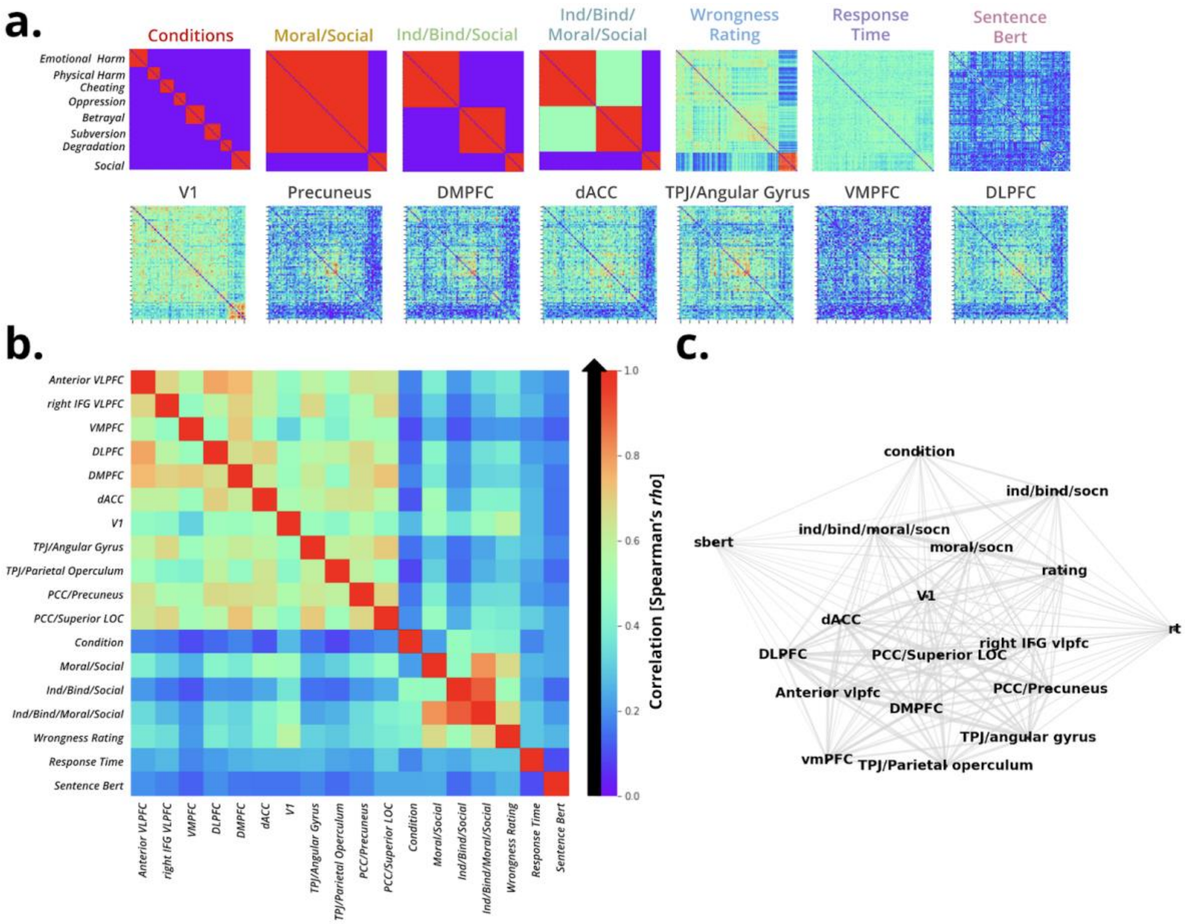


Figure 2. Data-driven visualization of relationships among multiple representational models. **a.** Representational similarity matrices (RSM) of conceptual, behavioral, embedding, and group-averaged brain representations across vignette items. **b.** Matrix of RSM correlations. Each entry compares two RSMs by Spearman's ρ. The correlation matrix is symmetric about a diagonal of ones and all correlations across RSMs were positive. **c**. Force-directed graph of the RSMs. Each point represents an RSM, and distances between the points approximate the ρ correlation distances among the RSMs.

This distinction is also reflected in individuals' behavioral wrongness ratings, highlighting that social norm transgressions, compared to moral norm transgressions, are judged more similarly. Compellingly, the close match between V1 and individuals' moral wrongness rating provides further evidence for the attentional capture (Gantman et al., 2020; Gantman & Van Bavel, 2014, 2015) and motivational role of visual imagery in intuitive moral judgments (Amit & Greene, 2012; Kahane et al., 2012). Yet, mounting evidence suggests that neural activity in V1 does not reflect the processes directly undergirding moral judgments, but the processes that trigger them by making aspects of a moral situation more salient. For example, the high similarity between representational patterns in V1 and dACC as well as dlPFC, two central ROIs in the moral judgment network, suggests that sensory representations about moral information are propagated throughout the brain. Additionally, when examining the correlations across RSMs (Figure 2b), more fine-grained subdimensions of moral actions come into focus. For example, V1, dACC, dlPFC, and anterior vlPFC contain more information about categorical divisions between individualizing and binding moral foundations than vmPFC and TPJ as visualized by the closer/further distance of these ROIs to the ind/bind/moral/social RSM in the force-directed graph of RSM correlations (Figure 2c). Moreover, the force-directed layout efficiently pushed models with an overall low similarity to the periphery, suggesting that the categorical divisions of moral foundations, response times, as well as the sBERT model are suboptimal explanations of the neural RSMs.

Subsequent to these data-driven, exploratory analyses, we tested the statistical significance of each candidate model's relatedness to each of the neural reference models (Figure 3). To this end, we followed established guidelines (Nili et al., 2014) using Spearman's rank correlation coefficient as measure of between-model similarity and performed a one-sided signed-rank test across the single-subject RSM correlations. Although the resulting correlations are relatively small – which is to be expected when using coarse grained ROIs, rather than more fine-grained, smaller nodes and spheres – all coefficients are statistically significant, suggesting that each candidate model explained variance in the neural representations of moral foundations.

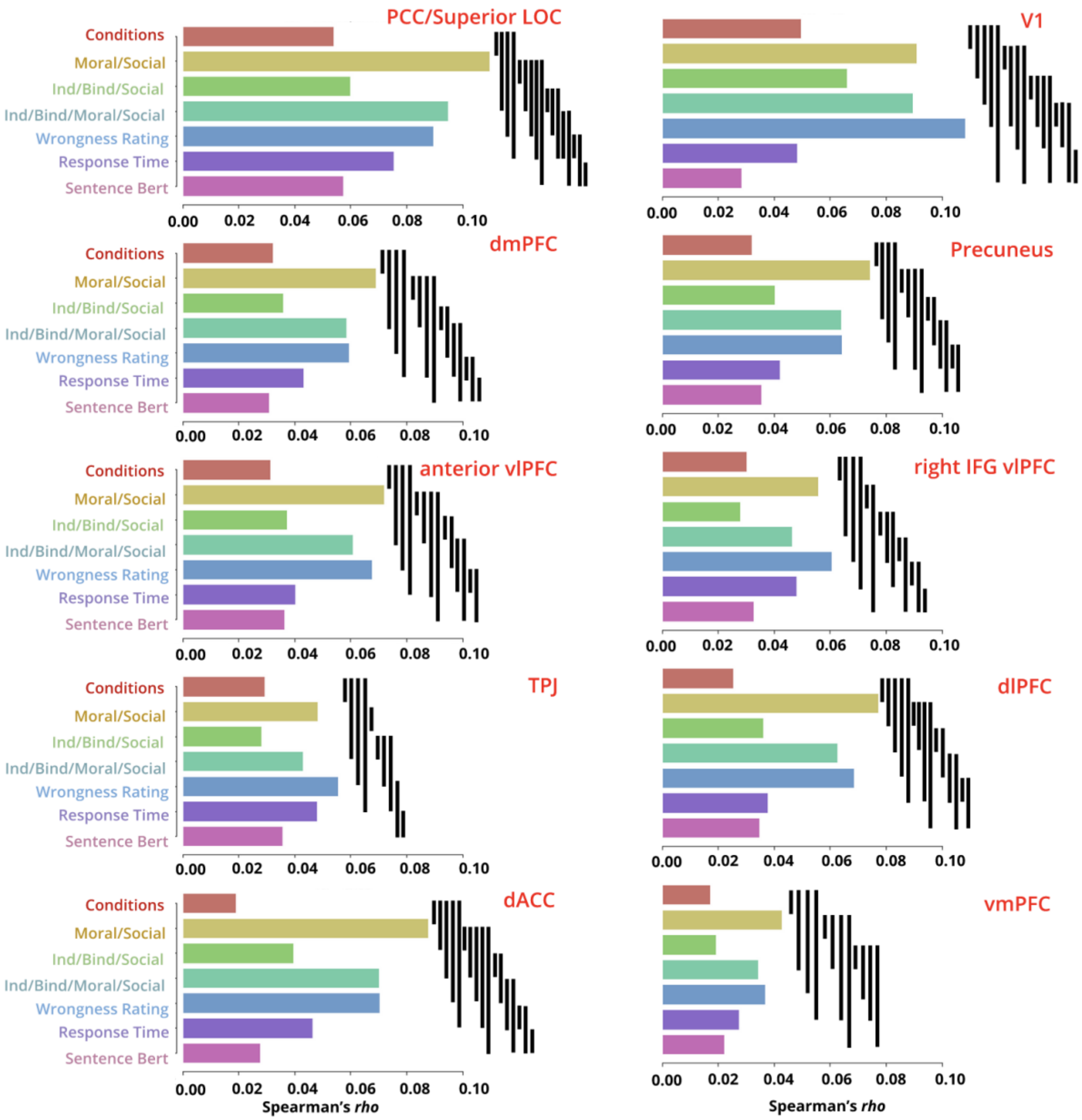


Figure 3. Inferential comparisons of model representations. Data-driven visualization of relationships among multiple representational models. Spearman's rank-order correlations (all $p < .001$) indicate the similarity between candidate model and neural pattern similarity. Black bars show significant pairwise differences between candidate models. Multiple testing was accounted for by controlling the FDR at $q < 0.05$.

To test whether two candidate models are statistically different in their relatedness to the neural representation, we computed the difference between the candidate model correlations in each subject and performed a two-sided signed-rank test across subjects. In the majority of ROIs, the conceptual moral/social model outperformed all other candidates, suggesting that the categorical division between moral versus social norm violations is salient within each ROI's neural activation pattern. The conceptual model defining individualizing versus binding foundations as a subcategory of the larger moral versus social model consistently outperformed the more limited individualizing/binding/social model. This indicates that the brain distinguishes between individualizing and binding foundations as two subcategories of a larger moral space whose activation pattern is distinct from social norm violations.

Intersubject Representational Similarity Analysis

Table 1. *Significant intersubject representational similarities for moral vignettes*

ID Measure	Model	Vignette Contrast	ROI	Spearman's *rho*
Empathy	Distress	carep	**dlPFC**	-0.082
	Fantasy	pur	?	0.187
	NN-Itemwise	carem	Temporal Occipital Fusiform	0.104
	NN-Itemwise	carep	V1	0.235
	NN-Itemwise	loy	Cerebellum (VI)	0.101
	NN-Itemwise	pur	V1	0.268
	Perspective T.	lib	**PCC/precuneus**	0.276

Table 1. *Contd.*

ID Measure	Model	Vignette Contrast	ROI	Spearman's *rho*
Moral Intuition Salience	AnnaK	loy	?	-0.236
	NN-Itemwise	auth	**dmPFC**	0.131
		carep	V1	0.228
		carep	?	0.128
		lib	?	0.103
		loy	anterior fronto-parietal (dlPFC, dACC)	0.171
		pur	?	0.142
Moral Wrongness Ratings	AnnaK	auth	**amPFC**	-0.160
		auth	**dACC**	-0.234
		carem	Amygdala	-0.124
		carep	**dACC**	-0.203
		lib	Amygdala	-0.123
		loy	**mid insula**	-0.267
		loy	Primary Auditory	-0.203
		pur	**anterior MPFC**	-0.230

Table 1. *Contd.*

ID Measure	Model	Vignette Contrast	ROI	Spearman's *rho*
Moral Wrongness Ratings	AnnaK	pur	**Dorsal Anterior Insula**	-0.192
		pur	?	-0.170
		pur	**dACC**	-0.311
		pur	Left IFG	-0.179
		pur	?	-0.171
		pur	?	-0.348
		pur	?	-0.280
	NN-Itemwise	auth	?	0.186
		carem	V1	0.255

Note. Condition Key: carep: physical harm; carem: Emotional harm; lib: Oppression; loy: Betrayal; auth: Subversion; pur: Degradation. All reported ROIs survived Bonferroni correction. '?' indicates a missing atlas label for brain parcellation. Bolded ROIs were functionally localized by the moral vignette localizer.

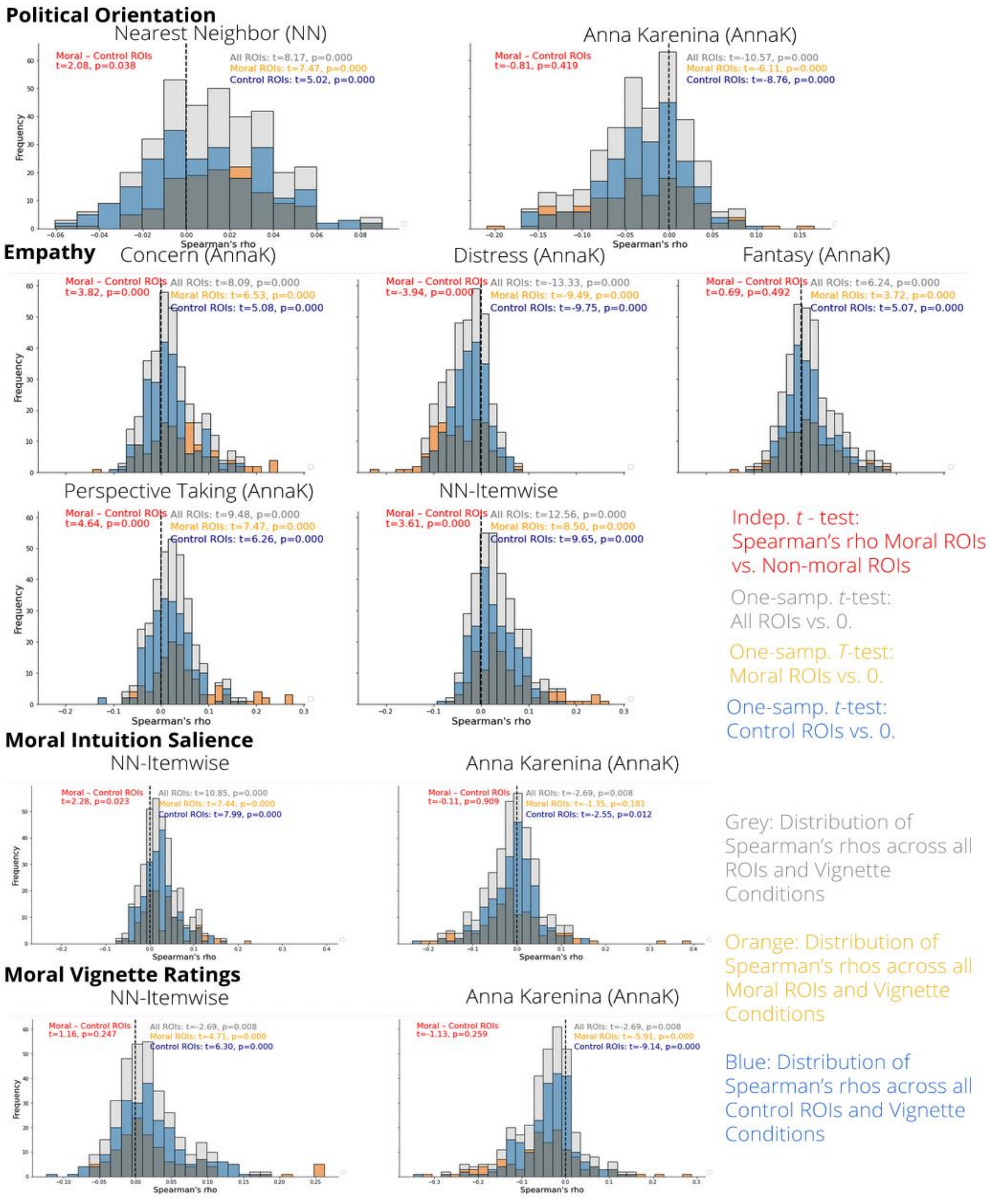


Figure 4. Distribution of Spearman's rho for moral vignettes per individual difference model

Narrated Movie Summaries

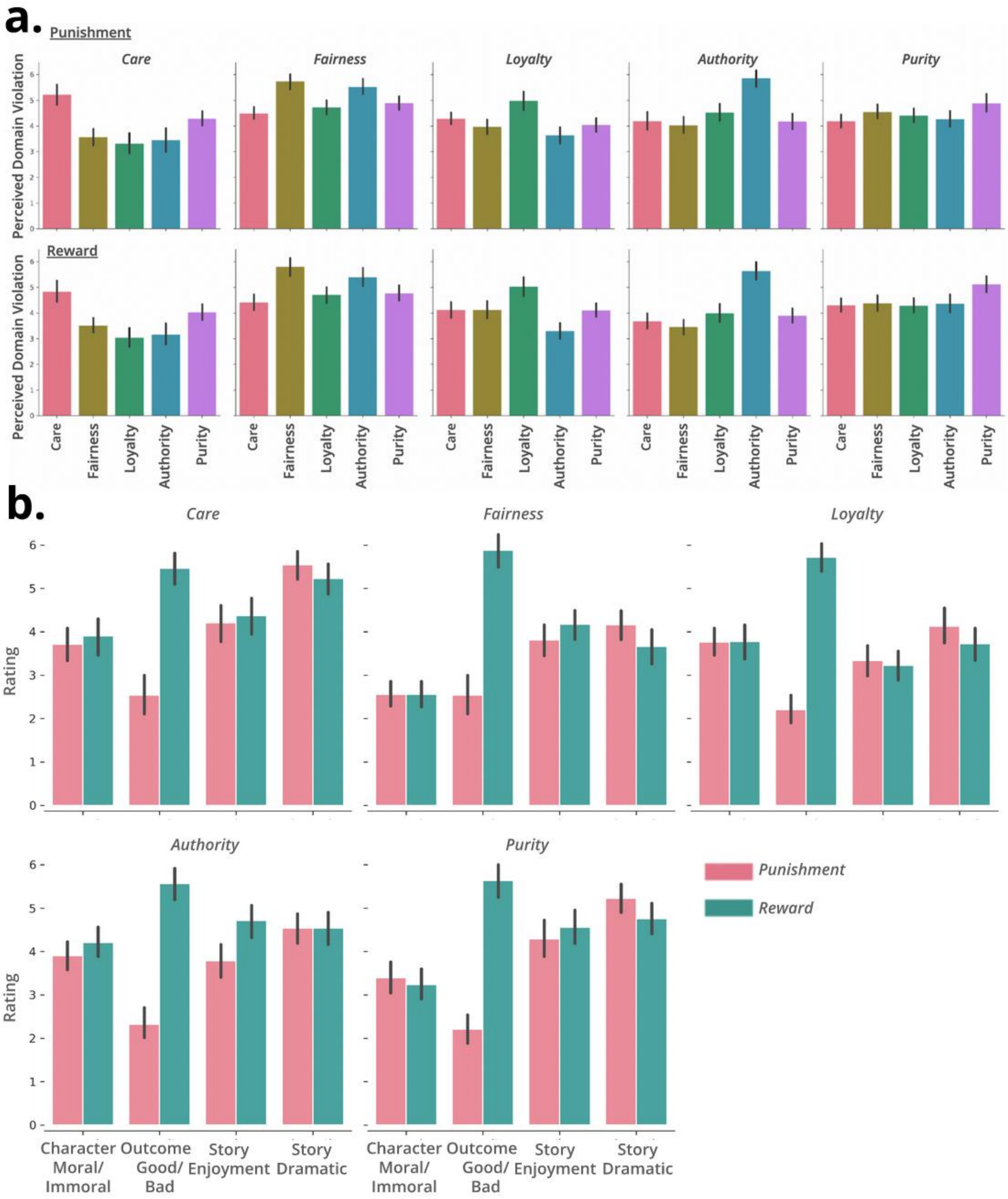


Figure 5. Behavioral results of narrated movie summaries. **a.** Perceived violation of moral foundations across stories. Each story violated a single moral foundation (see subtitle) while upholding all others. **b.** Effects of story outcome on perceived character morality, outcome valence, story enjoyment, and dramatic perception of story. Error bars reflect 95% confidence intervals based on 1,000 bootstrap iterations.

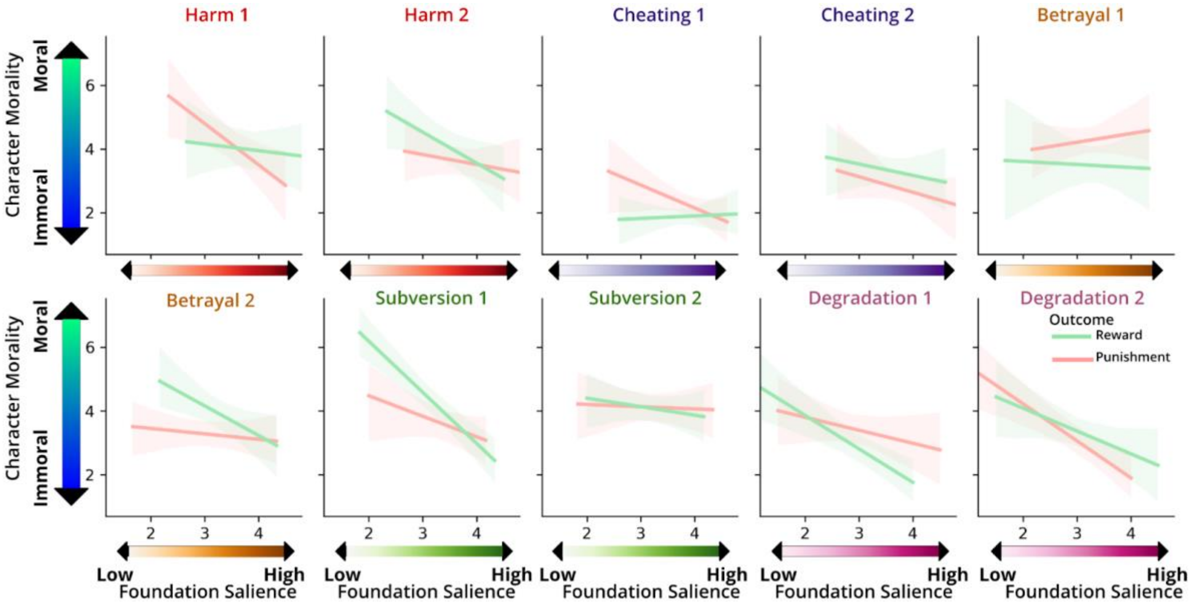


Figure 6. Moral foundation salience predicts rated character morality. For each story, the moral foundation salience score for the foundation that the story character violated (see title) served as predictor for rated character morality of that story. With the exception of two stories (betrayal 1; subversion 2), a higher foundation salience led to harsher (more negative) character judgments. Significant interactions based on character outcomes (red versus green graphs) did not emerge.

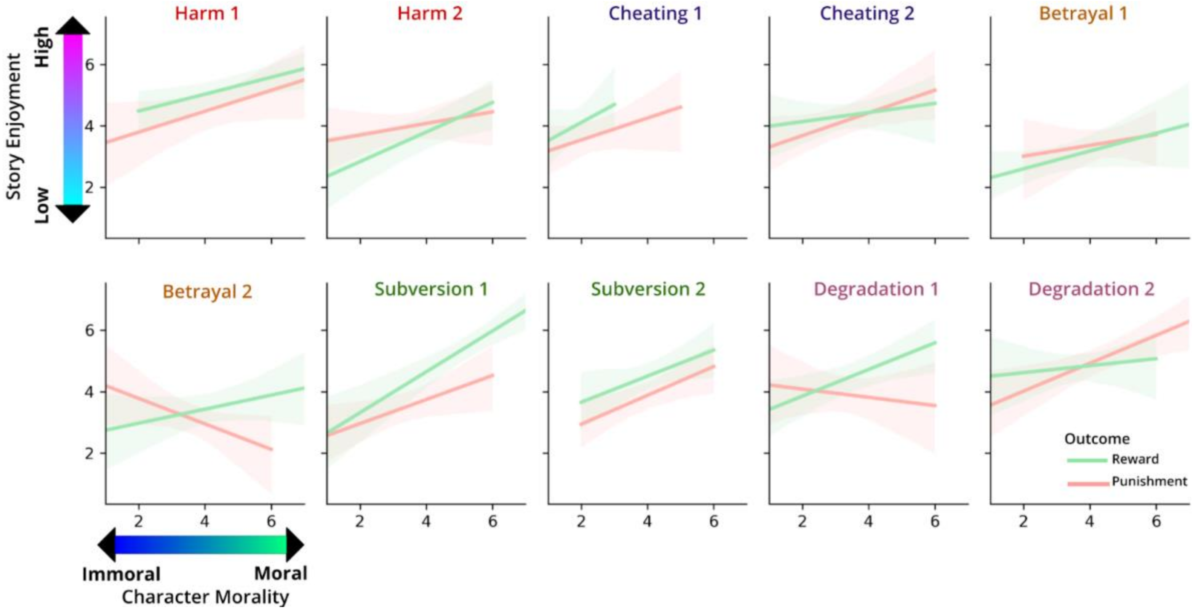


Figure 7. Character morality predicts story enjoyment. For each story, rated story enjoyment was regressed onto perceived character morality. We predicted that story enjoyment increases when a morally appraised character is rewarded or when an immorally appraised character is punished. In contrast, story enjoyment should decrease when an immorally appraised character is rewarded and a morally appraised character punished. While we do find that perceived character morality positively predicts story enjoyment, an interaction with character outcome was only apparent for two stories (betrayal 2; degradation 1).

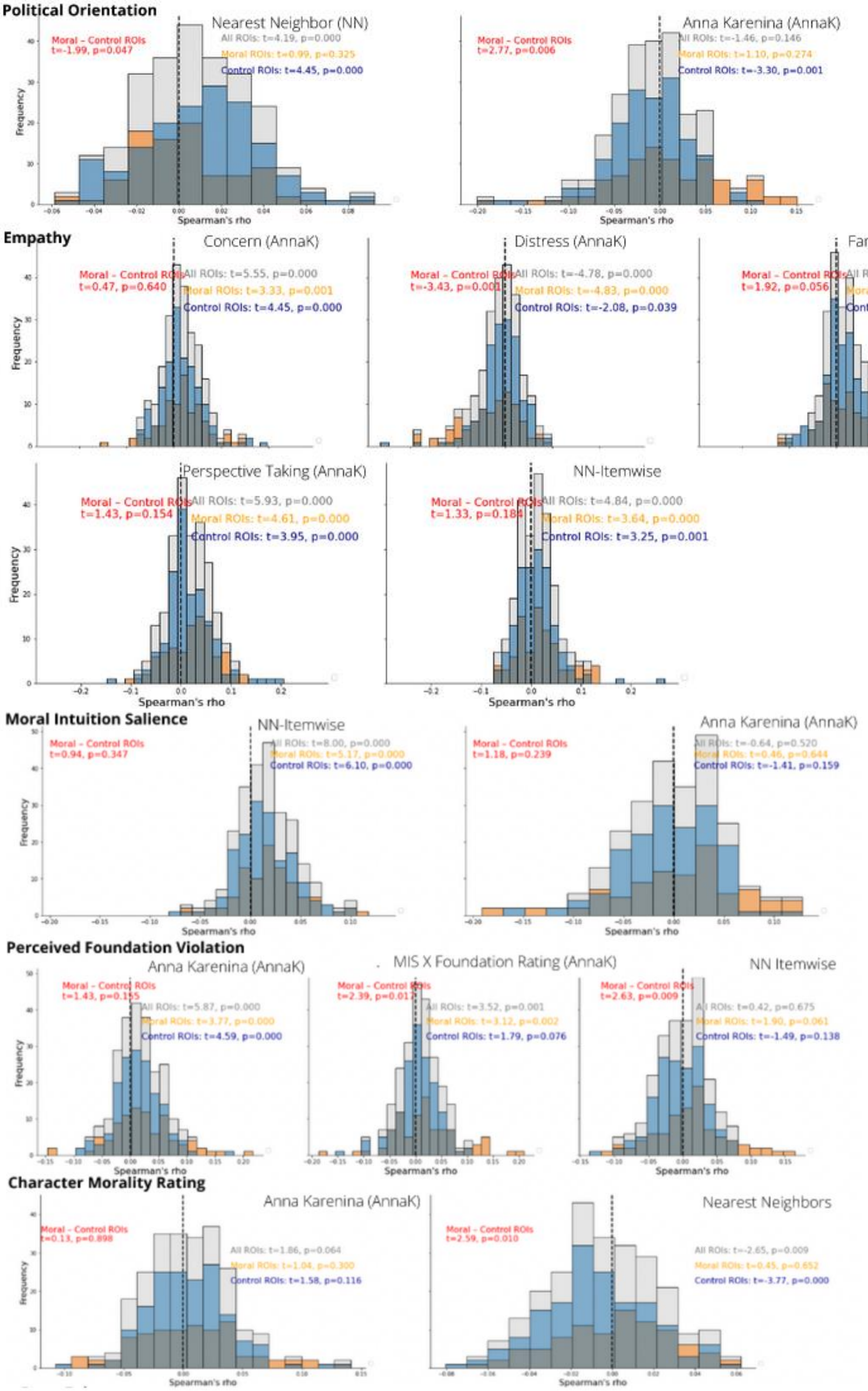


Figure 8. Distributions of Spearman's rho for movie summaries per individual difference model.

Political Attack Advertisements

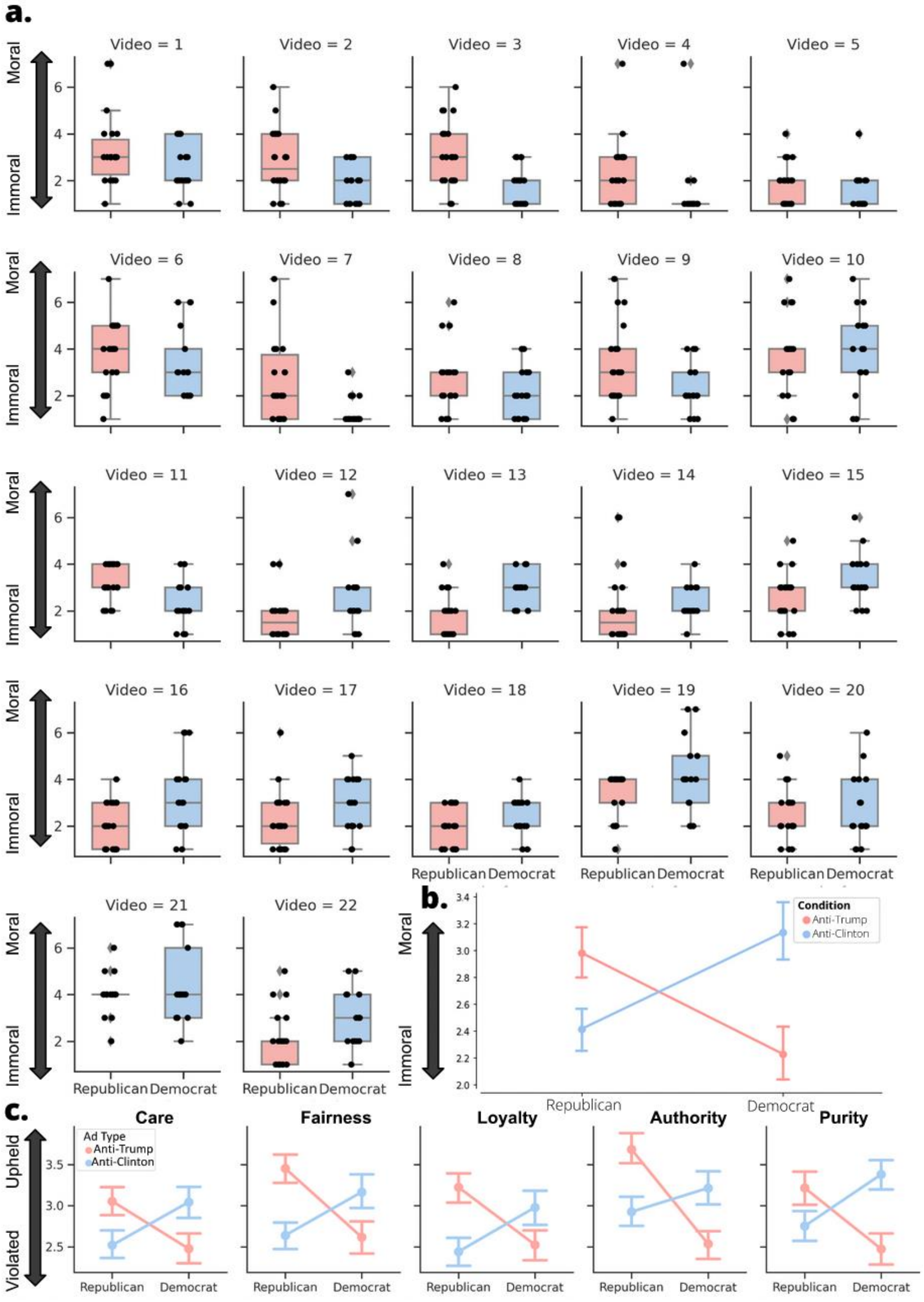


Figure 9. Political affiliation modulates moral judgment of political candidates. **a.** For each video, subjects rated the degree to which they perceived the candidate's behavior to be morally wrong. Videos 1–11 are anti-Trump; videos 12–22 are anti-Clinton. **b.** Republicans perceived Trump (Clinton) to act more morally (immorally) in anti-Trump (anti-Clinton) ads, whereas democrats perceived Clinton (Trump) to act more morally (immorally) in anti-Clinton (anti-Trump) ads **c.** Effects of **b.** persist for ratings of each moral foundation perceived to be violated (upheld) in each clip.

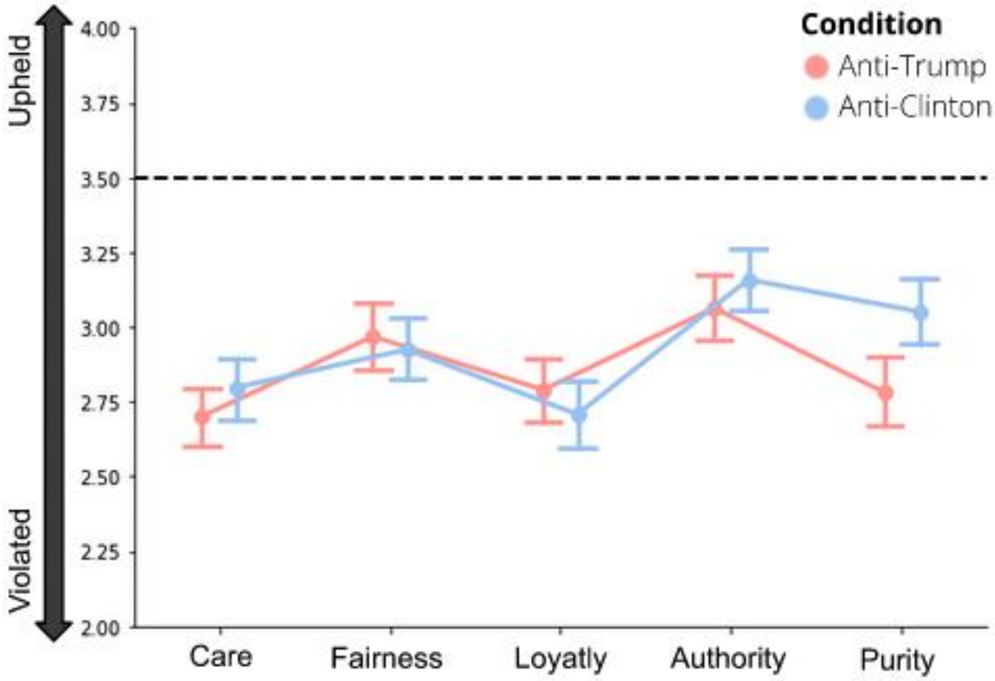


Figure 10. Perception of moral foundations for each condition across all participants. Individuals perceived Trump and Clinton to violate each moral foundation rather than to uphold them across conditions (mean of scale illustrated as dashed line). Significant differences only emerged for the purity foundation, where Trump was rated to violate purity more than Clinton.

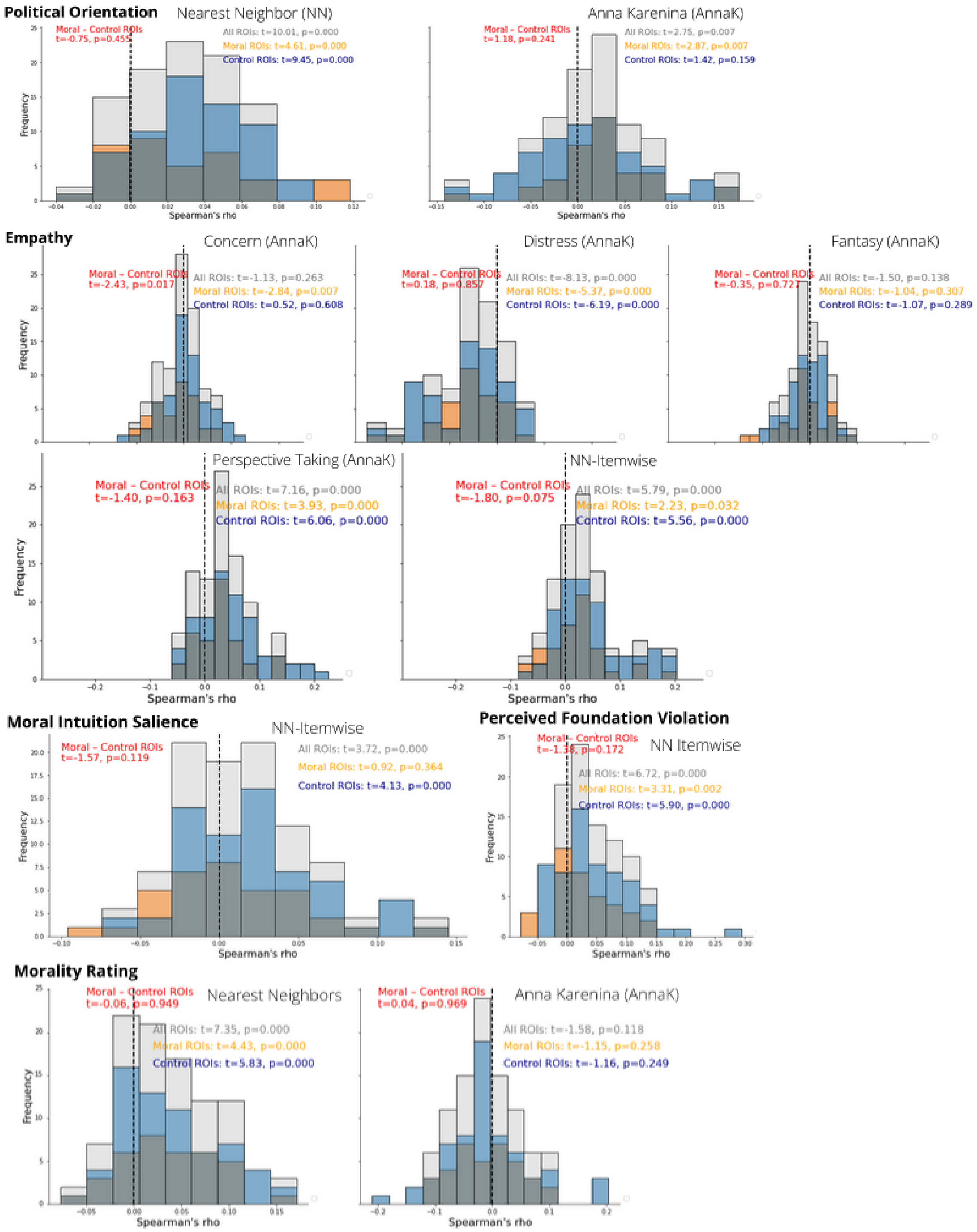


Figure 11. Distribution of Spearman’s rho for political ads per individual difference model.

Soap Opera Clips

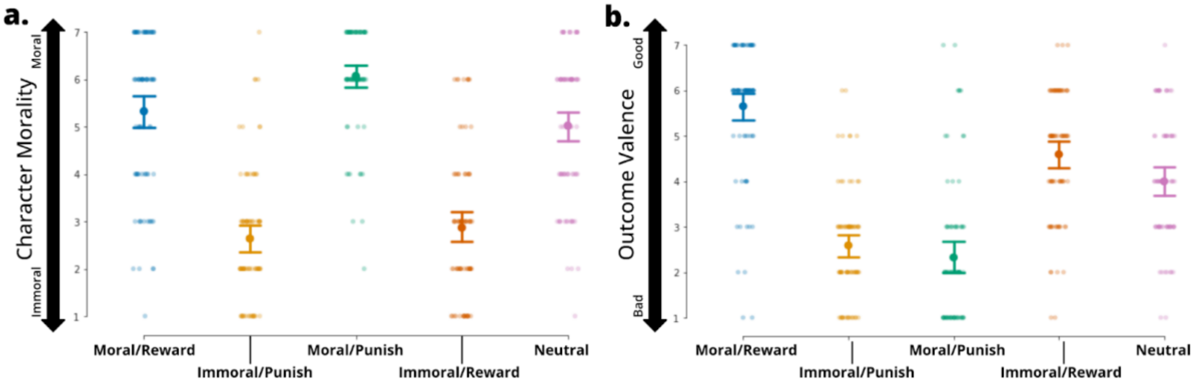


Figure 12. Distribution of Spearman's rho for political ads per individual difference model. Character morality and outcome valence ratings. **a.** The main character was rated as more moral in scenarios where the main character indeed acted more morally, whereas the main character was perceived as more immoral in scenarios that illustrated an immoral character. **b.** Outcomes for characters were rated as more positive if the scenario displayed a character that was rewarded, whereas outcomes for characters that were punished were appraised more negatively.

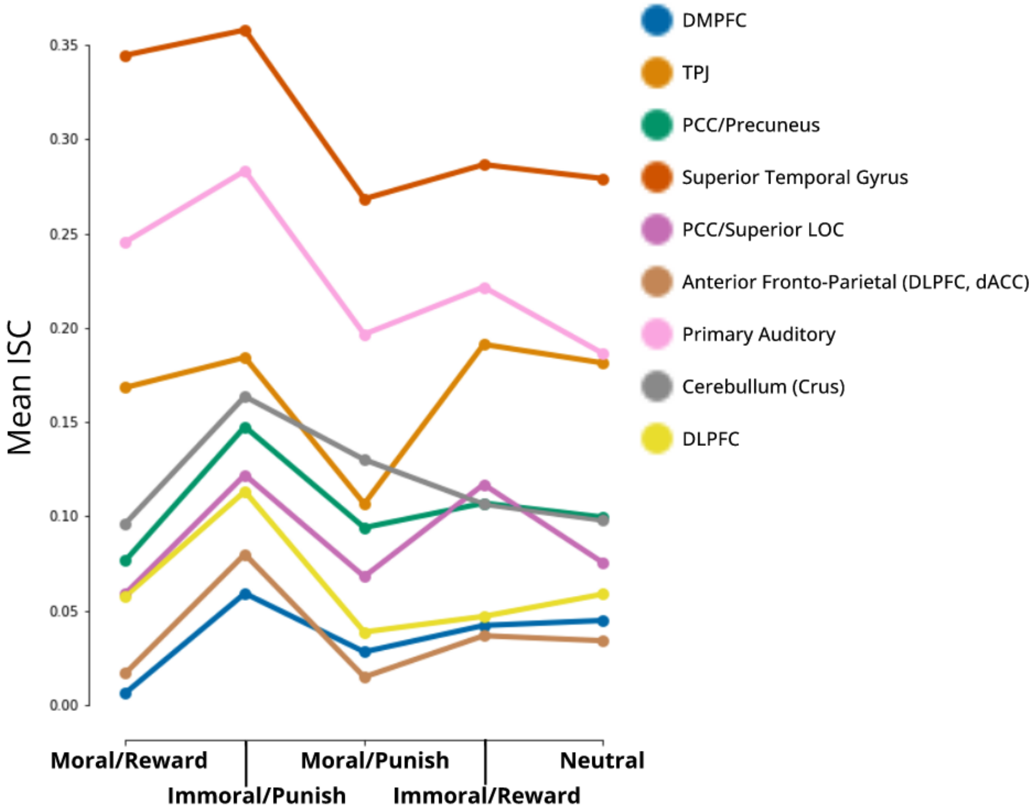


Figure 13. Comparison of intersubject correlation across moral ROIs and conditions.

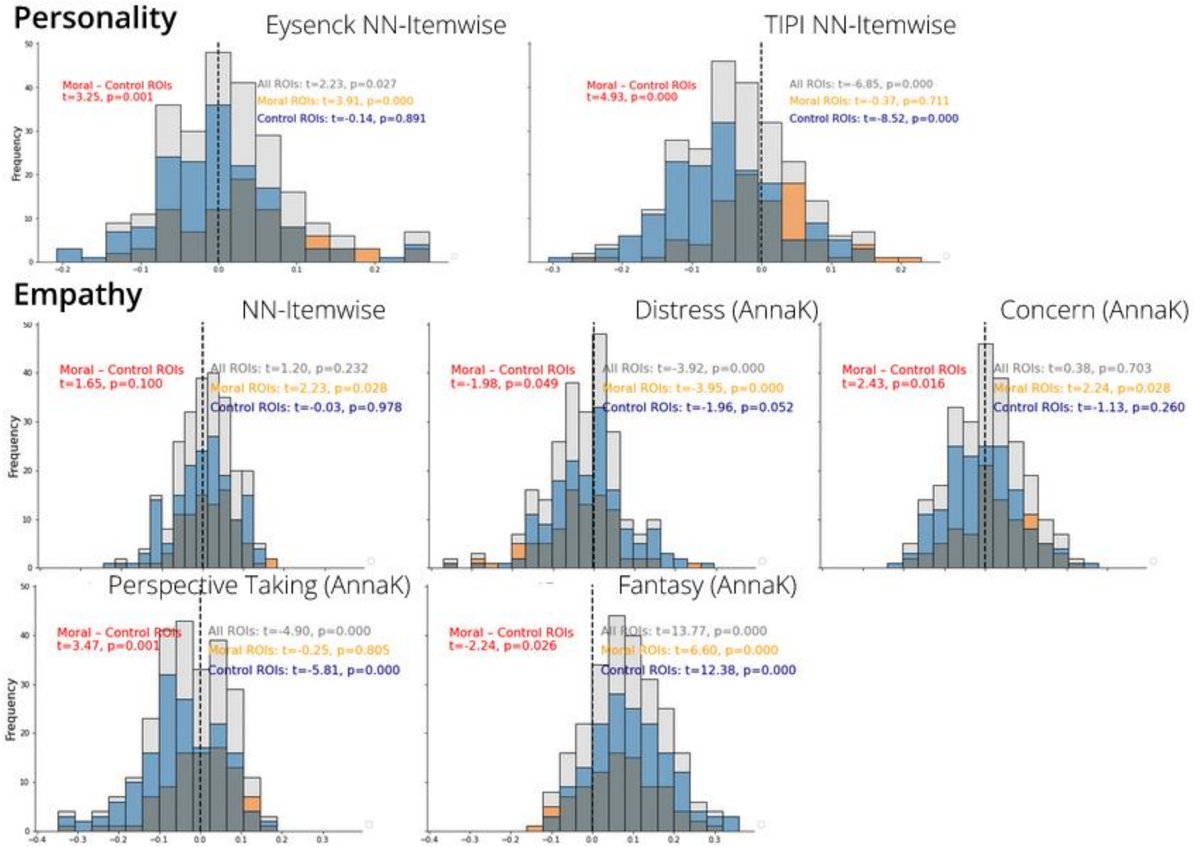


Figure 14. Distribution of Spearman's rho for soap opera clips per trait individual difference model.

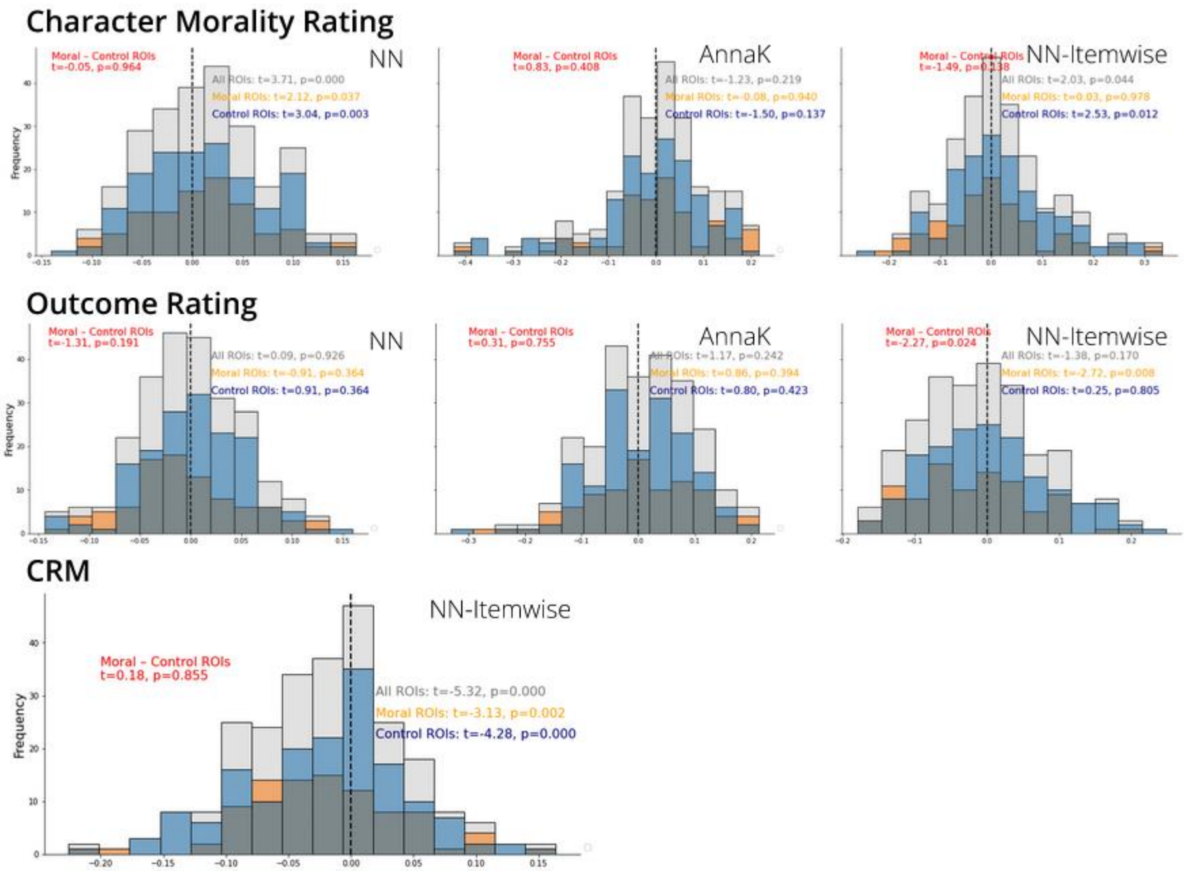


Figure 15. Distribution of Spearman’s rho for soap opera clips per behavioral individual difference model.

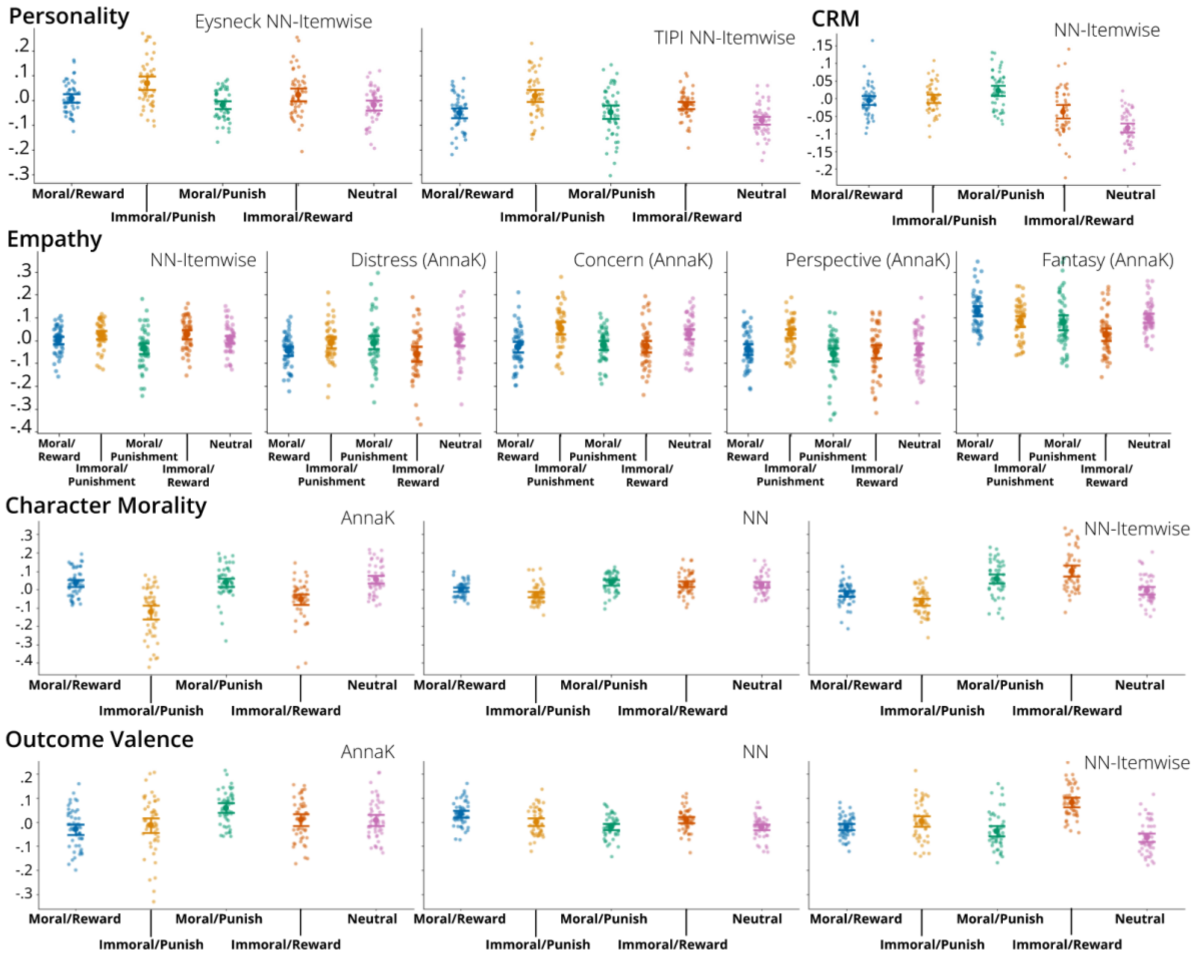


Figure 16. Intersubject representational similarity analysis for soap opera clips. Point- and stripplots for Spearman's rho across individual difference models and soap opera conditions.

Full Length Movie

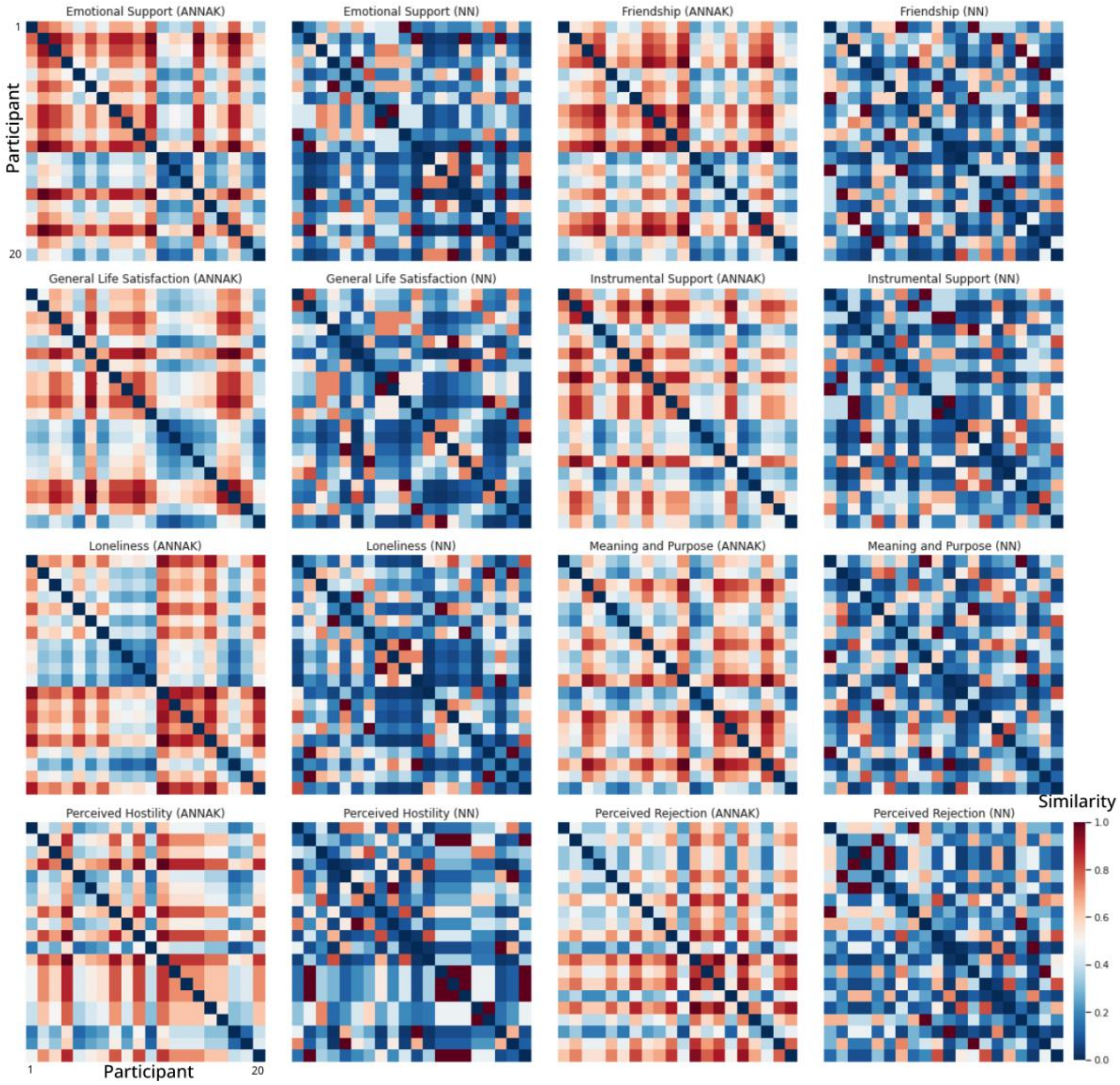


Figure 17. Similarity matrices for individual differences in NIH toolbox emotion measures.

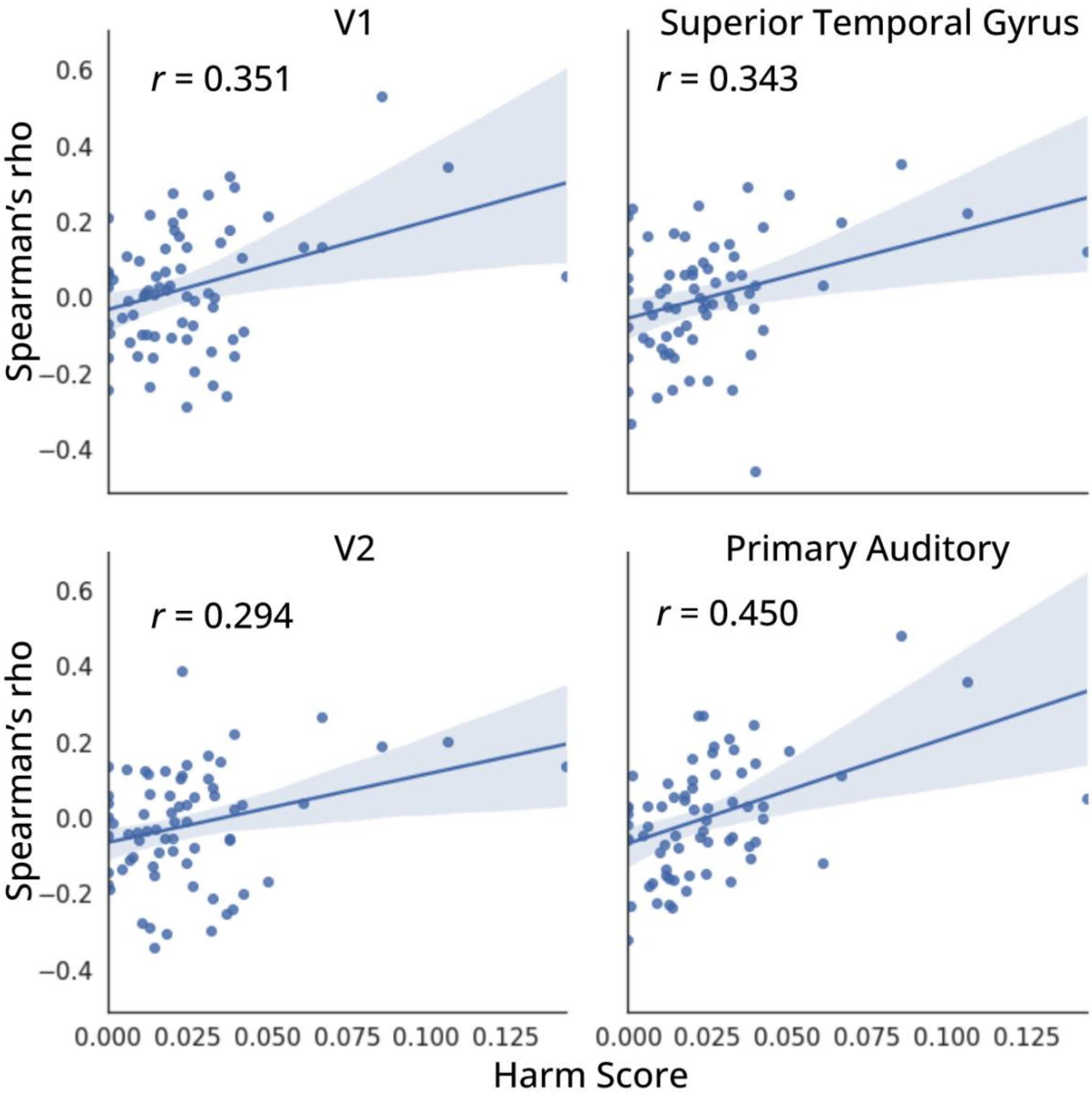


Figure 18. Harm cues increases intersubject neural synchrony among individuals high in perceived hostility.

Glossary

Anna Karenina (AnnaK): A type of similarity metric that captures how similar individuals are in *absolute* terms on an overall score of a given scale. Specifically, the AnnaK similarity structure tests if individuals who score high on a given scale are *similar* to other high-scorers and *dissimilar* to low-scorers. A powerful feature of operationalizing similarity in this way is that the same model can detect effects in both directions, based on the *sign* of the resulting Spearman's *rho* value between brain, trait, and behavioral similarity matrices. If high scorers are alike and low scorers different, the resulting Spearman's *rho* would be positive; if low scorers are alike and high scorers different, the resulting Spearman's *rho* would be negative.

Nearest Neighbor (NN): A type of similarity metric that captures how similar individuals are in *relative* terms on an overall score of a given scale. When relating similarity matrices of brains with an NN similarity model, a positive Spearman's *rho* highlights that subject who are more similar to their immediate neighbors on a given trait/behavior scale show more *similar* neural responses, whereas a negative resulting Spearman's *rho* would suggest that subject who are more similar to their immediate neighbors on a given trait/behavior scale show more *dissimilar* neural responses.

Nearest Neighbor Itemwise (NN-itemwise): A type of similarity metric that captures how similar individuals are in *relative* terms in their item to item responses to a questionnaire. When correlating neural similarity matrices with an NN-itemwise similarity matrix, a positive Spearman's *rho* highlights that subject who filled out a questionnaire in more similar ways show more *similar* neural responses, whereas a negative resulting Spearman's *rho* would suggest that subject who filled out a questionnaire in more similar ways show more *dissimilar* neural responses.