NeuroImage 159 (2017) 371–387

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis

E.A. Wasserman [a,*], A. Chakroff [a], R. Saxe [b], L. Young [a]

[a] Dept. of Psychology, Boston College, Chestnut Hill, MA, United States
[b] Dept. of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States

ARTICLE INFO

Keywords:
Moral psychology
Representational similarity analysis
fMRI
Social neuroscience

ABSTRACT

Characterizing how representations of moral violations are organized, cognitively and neurally, is central to understanding how people conceive and judge them. Past work has identified brain regions that represent morally relevant features and distinguish moral domains, but has not yet advanced a broader account of where and on what basis neural representations of moral violations are organized. With searchlight representational similarity analysis, we investigate where category membership drives similarity in neural patterns during moral judgment of violations from two key moral domains: Harm and Purity. Representations converge across domains in a network of regions resembling the mentalizing network. However, Harm and Purity violation representations respectively converge in different regions: precuneus (PC) and left inferior frontal gyrus (LIFG). Examining substructure within moral domains, Harm violations converge in PC regardless of subdomain (physical harms, psychological harms), while Purity subdomains (pathogen-related violations, sex-related violations) converge in distinct sets of regions – mirroring a dissociation observed in principal-component analysis of behavioral data. Further, we find initial evidence for representation of morally relevant features within these two domain-encoding regions. The present analyses offer a case study for understanding how organization within the complex conceptual space of moral violations is reflected in the organization of neural patterns across the cortex.

1. Introduction

Judging an act as "morally wrong" may subjectively feel easy and instinctive; yet, underlying each judgment may be a complex, feature-rich representation of the act committed. A wrong act may take many physical forms, from pushing a button to pushing a man off a bridge (Greene et al., 2009), from a mere spoken word (Helwig et al., 2001) to a violent stabbing (Cushman et al., 2012). The victim may be another person or the violator themselves (Chakroff et al., 2013). Moral judgments may demand mental state representations: was the actor internally or externally motivated (Chakroff and Young, 2015)? Did she do it on purpose (Young et al., 2007)? At a higher level, the act may be represented as an instance of a more abstract conceptual category, such as 'harm-based' or 'purity-based' violations, and judged accordingly (Graham et al., 2012; Dungan and Young, 2012; Chakroff et al., 2016b).

Understanding the organization of these representations is critical to understanding how humans conceive of and reason about morally charged acts. Indeed, a long tradition of moral psychological work has sought to answer questions of organization: on what basis can moral acts be grouped? Turiel's classic Domain Theory sought to draw a boundary separating *morals* from *conventions*, on the grounds that morals are generalizable: a moral violation is wrong everywhere and always, even if it is socially condoned (Turiel, 1983). Moreover, moral violations are intrinsically *harmful*, unlike norm violations, which may be merely awkward or improper. With a similar goal, Nichols (2002) separates moral from conventional by arguing that morals are "norms with feeling", defining moral violations as conventional violations accompanied by an affective response. Beyond circumscribing the moral sphere, the problem of organizing morals *within* the sphere has been addressed by Moral Foundations Theory (Haidt et al., 1993; Graham et al., 2012), which argues that morals fall into five principal domains, each characterized by a specific value and its antithesis (loyalty/disloyalty, fairness/cheating, authority/rebellion, purity/impurity, or care/harm).

To translate this question of structure among moral representations into the neural realm, we reframe it in terms of hypotheses about two basic organizing principles: similarity and hierarchy. *Similarity* among representations can reveal basic clustering structure within the space of violations, while assessing *hierarchy* can illuminate how the mind nests

* Corresponding author.
E-mail address: emily.wasserman@bc.edu (E.A. Wasserman).

http://dx.doi.org/10.1016/j.neuroimage.2017.07.043
Received 30 March 2017; Received in revised form 26 June 2017; Accepted 19 July 2017
Available online 22 July 2017

similarity-based clusters to achieve balance between structural parsimony and complexity. We use searchlight representational similarity analysis (RSA) to test a particular model of organization, based on a two-domain model derived from past work (Dungan and Young, 2012; Chakroff, 2015; Chakroff et al., 2016a, 2016b), as a case study to investigate how experimentally determined similarity and hierarchy manifest in converging neural representations across the cortex. Further, in exploratory analyses, we examine representational similarity based on a limited set of psychologically plausible features, as a first effort to determine whether morally relevant features are also being represented in the cortical areas most responsible for representing moral-violation concepts.

As in much RSA work, we employ stimuli that have been structured *a priori*, into two moral domains (Harm and Purity) and four moral subdomains. This method may be seen as analogous to the use of supervised learning models (versus unsupervised models) in data analysis. While we cannot directly assess how the brain *naturally* organizes its representations when encountering unstructured sets of violations, we can assess whether and where it is able to replicate a predefined organizational structure.

1.1. Neural representations of violations

Previous neuroscientific work on morality has largely addressed questions of *content* – where morally relevant features are processed – rather than *structure*. For example, this work has found that the ventromedial prefrontal cortex represents social-emotional value for moral judgment (Koenigs et al., 2007; Shenhav and Greene, 2014) and that the right temporoparietal junction (RTPJ) represents and integrates mental state information for moral judgment (Young and Saxe, 2008; Young et al., 2007). Different affective responses to violations – e.g., moral disgust elicited by impure acts versus indignation elicited by harmful acts – are reflected in BOLD activation differences in various brain regions, including bilateral inferior frontal gyri (Moll et al., 2002). To the extent that this work examines structure, it has taken a univariate functional-mapping approach, identifying regions that respond preferentially to violations of a certain type to argue for the functional coherence of certain groups of moral violations. The impure versus harmful distinction mentioned above, when framed as a distinction between the conceptual domains of Purity and Harm themselves rather than between their associated affective states, is reflected in BOLD differences in whole brain and region of interest (ROI) analyses (Parkinson et al., 2011; Borg et al., 2008; Chakroff et al., 2016a).

This approach answers a useful question – which regions are engaged more during the processing of a given violation type – but does not address the question of which regions, if any, show convergence of multivoxel patterns for violations of that type. Theoretically, pattern representations of a certain type of violation could all resemble one another in a given region without that region showing any preferential BOLD response to those violations, and conversely, a higher BOLD signal does not guarantee similarity of the underlying patterns. More recent work has taken a first step toward representational similarity hypotheses by investigating how morally relevant distinctions are reflected in multivariate pattern differences within neural regions. For example, multivoxel pattern classifiers (MVPA) have identified a binary intentional-accidental distinction in RTPJ's voxel patterns (Koster-Hale et al., 2013; Chakroff et al., 2016a), implying some degree of representational similarity within each violation type. Yet a comprehensive account of how moral-violation pattern representations converge differentially across the whole brain – a cortical map of moral-conceptual organization – remains to be discovered.

In other domains, the representational similarity approach has been highly successful in revealing cognitive organization across broad areas of cortex by characterizing the relationships between multivariate neural representations (Kriegeskorte et al., 2008; Davis and Poldrack, 2014). RSA and related methods have been fruitful in characterizing the structure of the space of physical object representations (Kriegeskorte et al., 2008), semantic representations (Handjaras et al., 2016; Huth et al., 2012), and lexical representations (Su et al., 2012) – as well as the key features driving structural organization. Yet their application to conceptual spaces involving social content is so far limited. For example, RSA has been employed to uncover dimensions of social-information representation within the mentalizing network (Tamir et al., 2015; Chavez and Heatherton, 2015) and belief attributions across the cortex (Leshinskaya et al., 2017). The moral representations tested here, as a subclass of social representations, thus present a novel challenge and opportunity for representational similarity analysis. If representational similarity can shed light on the neural and cognitive organization of objects, words, and concepts, can it do the same for moral violations?

2. Method

2.1. Participants (fMRI)

Forty-five adults participated in the study for payment. Six were excluded for missing or improperly recorded data, for a total sample size of 39 ($N = 10$ female), mean age 30.33 years. Of these, 14 ($N = 2$ female) were diagnosed with Autism Spectrum Disorder by a licensed clinician, based on Autism Quotient (AQ) scores. No group differences in RSA maps were found (see Supplementary Materials). All participants were right-handed native English speakers with normal or corrected-to-normal vision, and gave informed consent in line with institutional review procedure at MIT. Subsets of the data collected for this study have been previously reported in two published articles (Koster-Hale et al., 2013; Chakroff et al., 2016a); the sample reported here constitutes the full set of complete data available at the time of analysis.

2.2. Experimental design (fMRI)

Stimuli for the moral judgment task consisted of 60 written scenarios, of which 48 were moral-violation scenarios and 12 neutral social scenarios (for the full text of all scenarios, see Appendix A of the Supplementary Material). Within the moral scenarios, 24 depicted harm-domain violations, of which 12 were physical (e.g., poisoning) and 12 psychological (e.g., insults) violations. The other 24 depicted purity-domain violations, of which 12 were pathogen-based (e.g., drinking human blood) and 12 incest-based (e.g., consensual sex with an adult sibling) violations. Our choice of these two particular domains, as opposed to the five- or seven-domain Moral Foundations framework (Haidt et al., 1993; Graham et al., 2012), was motivated by the large body of existing literature that focuses on the harm-purity distinction in both psychological and neural responses (e.g., Chakroff and Young, 2015; Parkinson et al., 2011), and by our own past work suggesting that a two-type model captures most variation across moral judgments of actions (Dungan and Young, 2012). Each participant viewed all 60 scenarios in pseudorandom order across 6 runs, with condition order counterbalanced across runs and participants; no condition was shown twice in a row.

Each scenario was split into four serially presented segments - Background (6 s), Action (4 s), Outcome (4 s), and Intent (4 s; Fig. 1). In a subsequent 4-s window, participants judged the moral wrongness of the scenario on a scale from 1 (“not at all morally wrong”) to 4 (“very morally wrong”) using a button box. In the Intent segment, information was presented which either specified that the act was committed intentionally, with full knowledge (e.g., you knew that your sexual partner was your sibling and decided to commit incest anyway), or that the act was committed accidentally, in ignorance (e.g., your sexual partner was a long-lost sibling you didn't recognize). Intent was described with three categories of mental-state verbs: knowledge (knew/thought), realization (realized/discovered), and perception (saw/noticed). Half of the scenarios were randomly presented as intentional and half as accidental. No participant saw both versions of the same scenario.