Title: Social cognitive processes explain bias in juror decisions

Authors:

Jaime J. Castrellon[1,2], Shabnam Hakimi[2,3], Jacob M. Parelman[3,4], Lun Yin[2], Jonathan R. Law[1,2], Jesse A.G. Skene[1,2], David A. Ball[5], Artemis Malekpour[5], Donald H. Beskind[6], Neil Vidmar[6], John M. Pearson[1,2,7,8], J. H. Pate Skene[3,8,9]*, R. McKell Carter[3,10,11]*

* Senior authors contributed equally to this work.

Affiliations:

1 Department of Psychology and Neuroscience, Duke University

2 Center for Cognitive Neuroscience, Duke University

3 Institute of Cognitive Science, University of Colorado Boulder

4 Annenberg School for Communication, University of Pennsylvania

5 Malekpour & Ball Consulting (JuryWatch, Inc.)

6 School of Law, Duke University

7 Departments of Biostatistics & Bioinformatics, Duke University

8 Department of Neurobiology, Duke University

9 Initiative in Science and Society, Duke University

10 Department of Psychology and Neuroscience, University of Colorado Boulder

11 Electrical, Computer, & Energy Engineering, University of Colorado Boulder

Acknowledgements: We thank Tal Yarkoni for his assistance in accessing the Neurosynth database. This work was supported by: Duke Institute for Brain Sciences, ICS at CU Boulder, ONR via eCortex, N00014-18-C-2067 to R.M.C, NSF 1655445 to J.H.P.S, NIH K01-ES-025442 to J.M.P., and NSF DGE-1644868 to J.J.C.. Funders had no role in the study or manuscript. S.H. is currently at the Toyota Research Institute in Los Altos, CA.

Contributions: J.H.P.S., J.M.Pe, and R.M.C. conceived the project. All authors contributed to the task design. D.A.B., A.M., D.H.B., J.A.G.S, N.V., J.M.Pe, J.H.P.S., and R.M.C. wrote the crime scenarios and evidence modules. J.M.Pa, S.H., and R.M.C

designed and coded the fMRI task, and J.H.P.S. ran the fMRI experiments. S.H., J.M.Pa., and J.J.C. processed the data. J.M.Pe. and L.Y. wrote/designed the computational models of behavior. J.J.C., J.M.Pe, J.H.P.S, and R.M.C. wrote/designed the computational models of fMRI data. J.J.C. analyzed the data. J.J.C., J.H.P.S, and R.M.C. wrote the paper. All authors approved the final version of the manuscript.

Data availability:

Unthresholded fMRI statistical maps can be viewed/downloaded from Neurovault: https://neurovault.org/collections/11105/. Code and fMRI analysis pre-registration for hypothesis grouping can be viewed and downloaded from OSF: https://osf.io/rk92x/ and Github: https://github.com/jcastrel/juror_fmri_bias

Conflicts of interest:

The authors have no conflicts of interest to report.

Keywords:

Decision making, law, social bias, fMRI

Abstract: Jury decisions are among the most consequential social decisions in which bias plays a notable role. While courts take measures to reduce the influence of non-evidentiary factors, jurors may still incorporate biases into their decisions. One common bias, crime-type bias, is the extent to which the perceived strength of a prosecutor's case depends on the severity of the crime. Moral judgment, affect, and social cognition have been proposed as core processes underlying this and other biases. Behavioral evidence alone has been insufficient to distinguish these explanations. To identify the mechanism underlying crime-type bias, we collected fMRI patterns of brain activation from mock jurors reading criminal scenarios. Brain patterns from crime-type bias were most similar to those associated with social cognition (mentalizing and racial bias) but not affect or moral judgment. Our results support a central role for social cognition in juror decisions and suggest crime-type bias and cultural bias may arise from similar mechanisms.

Introduction:

Recent public discourse has heightened interest in the biases that make their way into complex social decisions. From public policy choices to decisions about vaccination or climate change, the way people interpret factual evidence has been shown to depend on social and cultural beliefs and attitudes (Unsworth and Fielding, 2014; Agarwal et al., 2021; Ruiz and Bell, 2021), which may also be related to additional, general, forms of bias. Juror decisions about whether a defendant is guilty or not guilty are among the most consequential decisions in society. Like other complex social decisions, juror decisions require people to evaluate the cumulative weight of the available evidence in light of personal experience, emotions, and biases. In addition to weighing evidence directly relevant to a decision, jurors tend to consider factors not logically informative for deciding the facts of the case (Devine, 2012). We refer to those factors as extra-evidentiary. Various extra-evidentiary factors can bias the outcome of juror decisions.

The most prominently studied of these biasing factors include racial or cultural prejudices and stereotypes. For example, a number of studies have focused on the effects of race and gender on convictions in the criminal justice system (Anwar et al., 2012; Devine et al., 2016; Flanagan, 2018). Experimental evidence complements this research. Mock jurors' decisions about guilt, culpability, and sentencing are biased against outgroup ethnicity and socioeconomic status regardless of the severity of the crime (Mitchell et al., 2005; Esqueda et al., 2008).

Research on racial and cultural bias has framed theories of cognitive mechanisms underlying bias in terms of emotional, moral, and social cognitive processes. Public understanding of racial prejudice has often assumed that prejudice and stereotypes are driven primarily by emotional animus or moral judgment (e.g., *dominative racism* (Kovel, 1984)). However, people can behave in racially biased ways even when they deny feelings of racial animus or superiority. Social cognitive theories like Aversive racism theory (Dovidio and Gaertner, 2004) and Social identity theory (Stets and Burke, 2000; Ingriselli, 2014) characterize bias in behavior as a means to preserve in-group value in the (claimed) absence of emotion. Are these emotional, moral, and social cognitive mechanisms inferred from race and cultural bias unique, or do they apply to other types of decision making biases?

In criminal justice decisions, biases related to certain types of crime are of particular concern. Documented cases of wrongful conviction have most often involved very serious crimes like murder or rape. It is not clear whether this reflects a higher rate of wrongful conviction for serious crimes or simply a greater likelihood that wrongful convictions will be discovered in those cases. Supporting the former possibility, several studies indicate that the seriousness of crimes can contribute to bias in juror decisions (Vidmar, 2002, 2003; Gastwirth and Sinclair, 2004; Wiener et al., 2006; Gross et al.,

2014). For example, Vidmar found that jurors are predisposed toward a belief in the guilt of anyone accused of sexual abuse (Vidmar, 1997). In a previous study using simplified crime scenarios, we found a more general relationship between the seriousness of a crime and mock juror ratings of the strength of the prosecutor's case against a particular defendant, independent of the evidence presented (Pearson et al., 2018). We refer to this as crime-type bias.

While social processes are proposed to underlie many forms of bias, several theories have been proposed to explain crime-type bias specifically, including the roles of emotional responses (Holloway and Wiener, 2018) and moral judgment (Greene and Haidt, 2002). Each of the theories proposes different cognitive processes as the mechanisms for bias. Understanding the precise underlying cognitive processes can shed light on how different kinds of biases arise. Because there are a number of combinations of cognitive processes that could lead to the same behavioral outcome, it is difficult to distinguish these theories using behavior alone. Resolving these overlapping social-affective explanations requires a new approach.

Here, we use neural data to test for the involvement of cognitive processes associated with crime-type bias by using whole brain pattern analysis (Li et al., 2017) to compare patterns of brain activation associated with a range of cognitive processes to those observed during the processing of crime scenarios independent of evidence. We then synthesize results from this pattern analysis to link social-affective models about extra-evidentiary factors from the literature. To implement this approach, we collected functional magnetic resonance imaging (fMRI) data from jury-eligible participants as they viewed a series of simplified crime scenarios that vary in the type of crime and amount of evidence supporting guilt (Pearson et al., 2018). In real-world criminal proceedings, courts often suggest that eliminating graphic descriptions or inflammatory language is sufficient to avoid prejudicial effects of very serious crimes. In our scenarios, the crime descriptions lack graphic details or other contextual information often considered to evoke prejudices or strong emotional reactions. Instead, we are able to focus on bias inherent in the accusation of different types of crime. We characterize the cognitive processes underlying this crime-type bias.

Materials and Methods

Subjects, recruitment, and sampling

Thirty-three healthy, jury-eligible adults between the ages of 18 and 52 were recruited from the Durham, NC community and underwent a single-session fMRI scan. All participants were screened for significant health or neurological problems and had normal or corrected-to-normal vision, and all gave written informed consent before the start of the experiment. Neuroimaging data from four participants were excluded from

the analysis because of technical issues. Twenty-nine participants (mean age = 30.6 ± 9.58 years, 15 women and 14 men) remained for neuroimaging analyses. The experiment was approved by the local ethics committee at Duke University Medical Center.

Experimental procedures

The mock-juror task (Fig. 1A), answering a call to use engaging narratives in neuroscience (Willems et al., 2020), was adapted from earlier work (Pearson et al., 2018) for use in the fMRI scanner. Subjects viewed 33 scenarios, each of which describes an accusation of a specific crime — from shoplifting to rape, murder, and child sexual abuse. The task was split into three runs with 10-12 trials (scenarios) per run. For each scenario, participants first viewed a textual description of a crime type (a 'scenario'; text was presented one sentence at a time across multiple screens for 5-30s). The initial description contained no evidence of who committed the crime. While type of crime was the primary difference between scenarios, the scenarios also differed in other details, including names of defendants and victims, and the circumstances of the crime. These other details were varied to keep subjects engaged throughout the task and to encourage participants to treat each scenario as a distinct crime. At the same time, the varied descriptions raised the potential for differences other than crime type to influence effects of the crime scenario on judgments of guilt. However, the large number of scenarios and the range of crime seriousness across those scenarios provided an opportunity to test for a main effect of crime seriousness on judgments of guilt (for scenario and evidence text see Table S1).

After viewing the crime scenario description, participants were shown three types of evidence implicating the named suspect (text presented one sentence at a time across multiple screens for 2-12s). Subjects viewed evidence that was: (1) physical evidence, (2) eyewitness, and (3) criminal history. The order of presentation for these evidence types was randomized. Types of evidence linking the accused to the crime included (1) one of three options of physical evidence (no physical evidence, DNA evidence or non-DNA physical evidence such as fingerprint or ballistic evidence); (2) either of two levels of eyewitness (eyewitness or a non-eyewitness) or (3) one of three options of criminal history (no prior convictions, prior conviction for a related crime, or prior conviction for an unrelated crime). This resulted in 18 unique evidence combinations (3 × 2 × 3) for each crime scenario. Participants viewed all 33 crime scenarios paired with only one randomized combination of evidence. Although each participant saw only one randomized combination of evidence for each crime scenario, we can test each of the 33 crime scenarios with all 18 evidence combinations (594 unique combinations) using a computational modeling approach (See Supplementary Materials for modeling details). Following presentation of the scenario and evidence,

participants were asked to rate the strength of the case against the accused and recommend a degree of punishment (on a scale of 0-100) for 6s each.

The crimes shown varied in seriousness according to their legal classification (misdemeanor versus felony) and victim type (victimless, loss of property, and personal harm or injury). For example, while a misdemeanor victimless crime described operating a still without a license, a misdemeanor loss of property crime described larceny and a misdemeanor crime about personal harm described vehicular manslaughter. Among felony crimes shown, a victimless crime described money laundering, a property loss crime described breaking and entering, and a crime with a personal harm described first-degree murder. See Tables S1 and S2 for detailed crime scenario texts and classification for all 33 scenarios.

Computational modeling of juror behavior

Using this task (Fig. 1A), we previously showed that evidence and extra-evidentiary information about the crime independently influence participant judgments about the strength of the case against the accused (Pearson et al., 2018). We used a hierarchical Bayesian model to estimate the effects of crime type and evidence type (independent of each other) on case strength and punishment ratings (see Supplementary Materials). Since this model has the advantage of accounting for sparsely sampled data, we can estimate effects for all scenarios even though participants did not view all possible combinations of scenarios and evidence types. We previously applied this model to a large online sample of participants (Pearson et al., 2018) and show here that the model replicates case strength and punishment effects (Fig. S1). As in the online sample, computational modeling of ratings by the current participants also distinguished the effects of different types of crime on rated case strength (Fig. 1B). Participants' mean case strength ratings for all fixed evidence are presented in Fig. S2.

We had hypothesized that the seriousness of a crime might affect the likelihood of guilt. In our task, crime seriousness can be indexed by subject ratings of how much punishment the crime deserves (Pearson et al., 2018). We used PCA to explore the relationship between crime seriousness and contribution of the crime scenario to perceived case strength. The first component describing shared variance between the two (PC1) explains ~95% of the variance (Fig. 1C). Similarly, in the prior online sample, PC1 explained 69.5% of the variability (Fig. S3). We therefore utilize the slope of PC1 as a summary of the common process, our operational definition of crime-type bias. To check that there were no additional processes driving similarity between punishment and the scenario contribution to case strength, we added PC2 as an additional explanatory variable in our analysis.

Univariate fMRI analysis

Acquisition and preprocessing were performed as described in supplementary information. Statistical modeling of fMRI data was performed using FSL FEAT (v 6.0.0) (Smith et al., 2004; Woolrich et al., 2009). First-level analyses used FILM prewhitening for autocorrelation correction. Event variables and parametric regressors were convolved with a double-gamma hemodynamic response function. For each subject, a general linear model was fit to the data with event regressors for (1.) scenario viewing, (2.) evidence viewing, (3.) case strength rating, and (4.) punishment rating. To identify regions associated with modulation of BOLD signal with crime-type bias, we included two parametric regressors for loading scores on the principal components (crime-type bias/PC1 and PC2) from the PCA described above between case strength and punishment model weights. Evidence accumulation during evidence viewing was modeled with a parametric regressor as the within-trial successive cumulative sum of evidence weights for case strength with each additional piece of evidence. Additional parametric regressors included the scenario-level case strength model weights during the case strength rating phase and scenario-level punishment model weights during the punishment rating phase. The present paper focuses on the parametric effect of crime-type bias during the scenario viewing event. Second-level analyses averaged within-subjects' data across runs using a fixed-effects model. Higher-level analysis used a mixed-effects model (FLAME1) to combine data across subjects. Statistical maps were thresholded using a cluster-forming threshold with a height of $Z > 2.3$, and cluster-corrected significance threshold of $p < 0.05$.

Neurosynth similarity analysis

In order to evaluate support for different explanations of sources of crime-type bias, we compared the parametric crime-type bias map to patterns taken from a meta-analytic database, Neurosynth (Yarkoni et al., 2011). This meta-analytic database identifies brain activation patterns associated with defined terms and topics from the neuroimaging literature.

We first performed a data-driven spatial correlation analysis between the crime-type bias activation map and each and every topic in the database. We estimated the spatial voxelwise Pearson correlation between our parametric fMRI maps and each of all the possible 200 topic maps in the Neurosynth database (version 5, 200 topics release: https://neurosynth.org/analyses/topics/v5-topics-200/) (Yarkoni et al., 2011). We visualized these correlations using histograms for thresholded (Fig. 2B) and unthresholded (Fig. S6) fMRI maps. We refer to this analysis as "model-free" because the topics are neither selected nor grouped on the basis of any theories of bias. We note, however, that there are assumptions that stem from the underlying topic model (Poldrack et al., 2012).

In addition to the model-free correlation analysis, we grouped topics according to their hypothesized role in theories of bias. We refer to this analysis as "model-based". We therefore constructed a series of a priori models which consisted of groups of cognitive processes associated with the social-affective model families. We grouped meta-analytic association (reverse inference) fMRI statistical maps from the Neurosynth database that reflect cognitive features associated with distinctive models of juror decision making. To limit user degrees of freedom while defining model structure, we preregistered the list of topics associated with each model of decision making (details at https://osf.io/rk92x/) prior to model-free and model-based analyses (Table S3). After defining the models, we ran spatial linear regression using the topic maps as predictors of each model's similarity to our parametric statistical maps for evidence accumulation and crime-type bias using the form $Y(fMRI\ map) = \beta 0 + \beta_1(topic_1) + \beta_2(topic_2) + \ldots + \beta_i(topic_i) + \epsilon)$ for *i* topics included in each model (Fig. S5). This allows for topic maps to vary in their contribution to the overall model's similarity to the crime-type bias map. We used both thresholded and unthresholded z-statistic maps from our results and unthresholded z-statistic Neurosynth topic maps provided by the Neurosynth developer, Tal Yarkoni (since the viewable maps on the Neurosynth website are thresholded). For thresholded map analysis, we restricted inclusion to voxels that survived cluster-based thresholding for positive parametric effects of crime-type bias (Fig. 2). This prevents describing a cognitive process as being involved in a decision if it was suppressed rather than activated.

To identify meaningful associations, we adopted a non-parametric approach. In order to conclude that one model of juror bias explains the decision-making process, the topics associated with that model should explain more variance than a randomly chosen set of topics (see SI for details).

Post-hoc analysis of crime-type bias

Next, we examined the features of crime scenarios that best account for crime-type bias. In particular, we explored why brain regions involved in social cognition explain much of the variance in crime-type bias. One possibility is that crime type could be driven by inferences about seriousness based on societal definitions according to legal classification and may engage similar cognitive processes as those associated with culture and ideation bias. Alternatively (or in addition), the effect of crime type is driven by theory of mind or perspective-taking from the point of view of a victim. We used ROI analysis to test whether crime seriousness (according to statute) or the type of victim harm (victimless, loss of property, and injury or loss of life) better correlate with activations associated with crime-type bias (see SI).

Following this, we performed a model comparison analysis to identify whether crime-type bias is best predicted by victim type or legal classification as singular

predictors, independent complementary predictors, or interacting predictors. We performed a likelihood-ratio test (^{2}) and evaluated the BIC scores of each model compared to an intercept-only model.

Results

Brain regions associated with social cognition track crime-type bias

To identify the patterns of brain activation associated with crime-type bias, we modeled the neuroimaging data using a parametric regressor (see SI). We observed constellations of activations associated with crime-type bias (Fig. 2A, SI). In particular, positive activations included temporal parietal junction, posterior cingulate and precuneus, thalamus, midbrain, striatum, and inferior frontal gyrus and negative activations included the medial prefrontal cortex. These brain regions are associated with a range of cognitive processes. No significant clusters were associated with the second principal component of the PCA (PC2). The unthresholded statistical map for crime-type bias is accessible on Neurovault (https://neurovault.org/collections/11105/). To clarify the cognitive processes linked to crime-type bias, we employed a data-driven approach to ask which of all 200 possible topics in the Neurosynth database (version 5, 200 topics release: https://neurosynth.org/analyses/topics/v5-topics-200/) (Yarkoni et al., 2011) were most strongly associated with crime-type bias.

To do so, we compared the patterns of activation from crime-type bias to association (reverse inference) maps for all topic maps included in the Neurosynth database. We limited our analysis to significant clusters of positive activation. We calculated the Pearson correlation between the thresholded parametric crime-type-bias map and each of the 200 topic maps in Neurosynth (Yarkoni et al., 2011) (see SI). Higher correlations reflect greater similarity between the pattern of activation associated with crime-type bias and the neural signatures associated with a given Neurosynth topic. Notably, the topic most strongly correlated with crime-type bias is topic 100 (r = 0.233) (Fig. 2B). Neurosynth terms associated with topic 100 include “race”, “racial”, “stereotypes”, “american”, “black”, “white”,“prejudice”, “biases”, and “chinese”, suggesting a possible link between the brain regions and cognitive processes associated with crime-type and racial biases. The top ten topics correlated with crime-type bias also includes several additional topics centered on social cognition,(topics 145 “mind, mental, social” topic: r = 0.213; 154 “social, interactions”: r = 0.126), or on brain regions regions often associated with social cognition (topic 123 “junction, tpj, temporoparietal”: r = 0.213; topic 155 “temporal, sulcus, superior”: r = 0.177) (Fig. 2B). The top ten Neurosynth topic maps correlated with crime-type bias did not include

topics related to affective processes or moral judgment. However, one topic related to moral judgment (topic 135 “moral”, “guilt”, “judgment”, “morality”, “justice”, “wrong”, “norms”, “good”, “legal”, “shame”) had a relatively high correlation with crime-type bias (r = 0.107) (Fig. 2B) although it was below the top 10 out of 200 topics. The magnitude of the correlations we observed here are within the range reported in other studies testing the similarity of Neurosynth topic maps with fMRI effects related to decision making (Bowring et al., 2019) and social cognition (Boccadoro et al., 2019).

Our model-free analysis suggested that crime-type bias might be associated with cognitive processes related to social cognition, to racial and other cultural biases, and potentially to moral judgment, but we did not find evidence that crime-type bias in our task is related to affect. To test those suggestions systematically, we selected and grouped topic maps from Neurosynth into four potential models (preregistered at https://osf.io/rk92x/): (1) affect, (2) moral judgment, (3) social cognition, and (4) culture and ideation bias (see Methods, Table S3). We compared the variance explained by each of our models to the null distribution for randomly assigned groups of topics randomly drawn from the 191 Neurosynth topics not associated with one of the a priori models. The null distribution for each test was defined for a number of topics matched to each model. Specifically, the affect model was compared to null models that contained 3 topics and the other models were compared to null models that contained 2 topics. Hereafter, a model is described as significant if it explains more variance than 95% of models composed of randomly drawn topics.

We find that patterns of brain activation in response to crime-type bias (Fig 2B) were strongly correlated with the social cognition model centered on theory of mind and mentalizing processes ($p = 0.0094$, resampled null, Fig. 3B) and the culture and ideation bias model centered on processing stereotypes and personal biases ($p = 0.0012$, resampled null, Fig. 3B). The affect (Fig. 3A) and moral judgment (Fig. 3B) models were not significant ($p = 0.802$ and $p = 0.331$ resampled nulls, respectively). Within the social cognition and culture and ideation bias models, three topics showed positive associations with crime-type bias: topic map 145 (“mind, mental, social”), topic map 154 (“social, interactions”), and topic map 100 (“gestures, abstract, race”).

Finally, we examined the specific brain regions that contribute to the similar activation patterns associated with crime-type bias, social cognition, and culture/ideation bias (Fig. 4A). The social cognition topic maps (145 and 154) overlap with activation for crime-type bias in the temporal parietal junction and posterior cingulate (Fig. 4A), which have been identified as hub regions for the mentalizing or theory-of-mind component of social cognition (Spreng and Andrews-Hanna, 2015). Notably, the overlap between crime-type bias and topic 100 also localized to the temporal parietal junction (Fig. 4A). This very specific overlap suggests that crime-type and racial biases may share at least one cognitive component related to social cognition.

Sensitivity to victim harm modulates crime-type bias

To test whether crime seriousness (according to legal classification) or victimhood was related to cognitive measures of crime-type bias (scenario PC1 from the model estimates of case strength and punishment), we applied an ANOVA for each classification type. This revealed a statistically-significant effect of victimhood ($F(2,30) = 32.08$, $p < 0.0001$) (Figure 4B) but not seriousness ($F(1,31) = 1.681$, $p = 0.204$) on PC1 (Figure S8A). Specifically, crimes resulting in injury or loss of life ($M_{PC1} = 2.345$, 95% CI = [1.607, 3.083]) were associated with substantially higher PC1 loading scores than victimless ($M_{PC1} = -1.26$, 95% CI = [-1.880, -0.637]) or property crimes ($M_{PC1} = 0.185$, 95% CI = [-0.643, 1.013]). To test whether crime seriousness or victimhood was related to neural measures of crime-type bias (mean fMRI activation from the ROI defined above), we again applied a mixed-effects model with a random intercept for each participant using the “lme4” and “lmerTest” packages in R. This revealed a statistically-significant effect of victimhood ($F(2,926) = 13.04$, $p < 0.0001$) (Figure 4C) but not seriousness ($F(1,927) = 3.25$, $p = 0.072$) (Figure S8B). Specifically, crimes resulting in injury or loss of life ($M_{z\text{-stat}} = 0.25$, 95% CI = [0.162, 0.428]) were associated with substantially increased and greater fMRI activation than victimless ($M_{z\text{-stat}} = -0.050$, 95% CI = [-0.219, 0.130]) or property crimes ($M_{z\text{-stat}} = 0.007$, 95% CI = [-0.098, 0.201]). Overall, we found that those scenarios that result in bodily harm or death (but not victimless or property loss crimes) drove both the increases in case-strength ratings (Fig. 4B, SI) and neural activation associated with crime-type bias (Fig. 4C, SI).

Although these analyses failed to identify strong associations between variability in legal classification and crime-type bias, the effect sizes do indicate some contributing role for legal classification in driving ratings. To pinpoint exactly how victim type and legal definition could each contribute to crime-type bias, we performed a model comparison testing whether the best model is explained by these variables as singular, independent, or interacting. A likelihood-ratio test (^{2}) using the “lrtest” package in R revealed that legal classification alone did not explain crime-type bias ratings better than an intercept-only model ($^2(1) = 1.74$, $p = 0.187$) and that victim type as a singular predictor provided a better fit than legal definition alone ($^2(1) = 35.99$, $p < 0.001$). Nevertheless, a model that included legal classification and victim type as independent predictors provided a better fit than a victim-only model ($^2(1) = 5.38$, $p = 0.020$). A model that included an interaction term between legal classification and victim type did not explain additional variance than the independent predictors ($^2(1) = 0.980$, $p = 0.613$). The BIC score difference (ΔBIC) from an intercept only model (which penalizes complex models) also indicated that the best parsimonious model included independent predictors for legal classification and victim type (Fig. S9). Similarly, fMRI activation associated with crime-type bias was also best explained by a linear model that included victim type as a singular predictor. Specifically, a likelihood-ratio test (^{2}) revealed that legal classification alone did not explain crime-type bias fMRI activation better than an

intercept-only model (^{2}(1) = 3.25, p = 0.072) and that victim type as a singular predictor provided a better fit than legal classification alone (^{2}(1) = 22.5, p < 0.001). Whereas a model that included legal classification and victim type as independent predictors did not provide a better fit than a victim-only model (^{2}(1) = 3.27, p = 0.071), a model that included an interaction term between legal classification and victim type did explain additional variance than the independent predictors (^{2}(2) = 10.65, p = 0.005). Since this interaction model is complex, the best parsimonious model according to the ΔBIC score was the victim-type only model (Fig. S9). Results from the model comparisons are shown in Table S8 and Table S9.

Discussion

We used the pattern of brain activations during juror decisions to show that crime-type bias is associated with social cognition. Notably, the cognitive processes we find to be associated with crime-type bias are similar to those that give rise to other biases like stereotypes about culture and race. Although bias is often thought to occur as part of an emotional response, we find that crime-type bias can occur even in the absence of strong emotional processes. It thus serves as a counter example to affect or animus as the sole driver of bias, one of the most popular explanations of bias.

Social cognitive processes track crime-type bias

Topic maps reflecting processes associated with culture and ideation bias and mentalizing are overlapping and highly similar to the parametric effect of crime-type bias (Fig. 4A), reflecting a common underlying use of social cognition. This is consistent with the theory of generic prejudice which speculates that jurors' personal knowledge, beliefs, and stereotypes can influence the perceived strength of the case against a particular defendant (Vidmar, 2003). In the context of this theory, our results indicate that crime-type bias may rely on similar neural and cognitive processes that give rise to other biases like stereotypes about culture and race.

Social cognition, and specifically the integration of culture, personal knowledge, and biases comprises a core component of broader models of narrative decision making (Shiller, 2017). In fact, Yuan et al. have noted the overlap between brain regions involved in narrative and social cognition (Yuan et al., 2018). This overlap is consistent with a proposed role for storytelling in the evolution of cooperation in human societies. Similar regions of the brain are also expanded in modern humans compared to non-human primates (Smaers and Vanier, 2019; Van Essen et al., 2019) and Neanderthals (Kochiyama et al., 2018; Neubauer et al., 2018). In this social role, narratives serve as means of transmitting cultural beliefs and knowledge beyond the first-hand experience of any individual (Smith et al., 2017; Bietti et al., 2019; Hitchcock, 2019).

In the context of juror decision making, narrative models propose that evidence integration by jurors relies on the extent to which the evidence can be assembled into a cohesive, compelling, and credible story (Pennington and Hastie, 1986; Shiller, 2017). Specifically, for jury decisions, narrative explanations are based on the extent to which the information about a case can be structured into a coherent account and how closely the competing narratives offered by prosecution and defense match the juror's background knowledge, experience, and beliefs (Arkes and Garske, 1982; Devine, 2012; Allen and Pardo, 2019; Hastie, 2019). Within this framework, social cognition could be understood as providing a basis for evaluating the plausibility of a narrative suggested by uncertain or incomplete evidence in a particular case.

We found that increased social cognitive processing was dependent on the presence of bodily harm as well as legal classification of the crime. This result is consistent with the notion that people rely on shared legal culture and norms in society that shape their understanding about how laws should be applied to others (Buckholtz and Marois, 2012; Vilares et al., 2017). These shared ideas about the law are infused in personal identities and inform narratives about society (Somers, 1994; Loseke, 2007) and are consistent with the social communication of laws and expectations through narratives.

While we have examined one specific type of bias that applies to juror decisions in criminal justice, the cognitive model we describe here suggests a more general context for understanding the interaction of biases, stereotypes, and cultural prejudices in complex decisions. In this view, bias does not necessarily require emotion or judgment, consistent with connections between social neuroscience and bias proposed by others (Amodio and Cikara, 2020). The overlap between neural correlates of social cognition and bias (noted also by others (Harris and Fiske, 2009; Cikara and Van Bavel, 2014)) implies that biases, stereotypes, and cultural beliefs are an important part of the knowledge base by which people make predictions about the world and evaluate alternative explanations, the narratives. The incorporation of social cognition in the narrative model provides a means by which bias (how likely is this story to happen in general?) and evidence presented at trial (what is the magnitude of support for this story?) can be integrated to reach a conclusion.

Relationship between crime-type bias and third-party punishment

Social cognition is considered critical to moral judgment from both theoretic (Gray et al., 2012; Royzman and Borislow, 2022) and cognitive neuroscience perspectives (Young et al., 2007). Notably, we did not observe an association between crime-type bias and moral judgment or affect (Kragel and LaBar, 2016). This is initially surprising because crime-type bias, as we operationalized it here, is correlated with subject ratings of the punishment deserved by each type of crime (Fig. 1C and Fig. S3). Previous studies have found that third-party punishment is correlated with activation of brain

regions associated with affect and norm enforcement, in some instances modulated by regions associated with social cognition (De Quervain et al., 2004; Buckholtz et al., 2008; Buckholtz and Marois, 2012; FeldmanHall et al., 2012; Treadway et al., 2014; Ginther et al., 2016; Stallen et al., 2018; Zinchenko, 2019).

The differences in brain activation between the two tasks emphasize that crime-type bias and punishment decisions represent separable processes. Punishment ratings reflect a direct assessment of crime seriousness, while crime-type bias represents an effect of crime seriousness on the probability it happened. Although the overall model for moral judgment was not associated with crime-type bias, one topic from that model (topic 135 - “moral”, “guilt”, “judgment”...) was relatively strongly correlated. Topic map 135 has a small overlap with crime-type bias activation within the TPJ and midbrain. For affect, brain regions involved in affective responses (e.g., amygdala) have been linked to moral judgments about actions that harm others (Treadway et al., 2014; Ngo et al., 2015) but results from our task show no direct link between affect and crime-type bias. This contrast emphasizes the difference between punishment and the decision about the strength of the prosecutor’s case.

The absence of affect as a component of crime-type bias is notable in the context of broader theories of bias (e.g. dominative racism). Our results are consistent with social-cognitive models of bias. It is important to point out that our experimental task, by design, presents a minimal description of the crime and the evidence, without graphic details or discussion of the harm to victims that might increase emotional or moral responses (see supplement). Our results therefore do not address the potential roles of affect and/or moral judgment on decisions about guilt in the context of a real trial. We are however, able to show significant bias in the absence of activation associated with affect. We therefore conclude that social cognition can drive bias in addition to any bias from affect and moral reactions. Even when steps are taken to mitigate emotional responses or moral outrage, bias driven by social cognitive processes may persist.

Limitations of the use of Neurosynth

We have attempted to separate the cognitive processes contributing to crime-type bias. In doing so, we have relied on a large body of literature from tens of thousands of researchers as well as a set of tools designed to assess the consistency of conclusions across those studies. While the use of meta-analytic data should increase reliability and generalizability, it also carries the limitations of the work on which it was based. We note that statistical maps from Neurosynth are based on association with a list of terms determined by a topic model generated by the authors of Neurosynth (Yarkoni et al., 2011; Poldrack et al., 2012). The use of a topic model then carries some assumptions: first, that the corpus of studies modeled in Neurosynth contains adequate representation of each of the targeted cognitive processes; second, that the researchers studied the intended concept; third, that the areas of research

treated as independent within the cognitive neuroscience community are indeed independent.

If these assumptions are violated, we would find a lack of association between the topics and fMRI activity included by Neurosynth. We therefore attempted to address these concerns by ensuring that each of our cognitive processes contained an adequate number of studies and showed clear clusters of significance in their maps. Each topic in the Neurosynth topic model includes at least 79 studies. The topics included in our categorical model comparisons each had significant clusters of associated activations. Within the model-free analysis, there are clearly topics for which there were not reliably associated clusters of activity. This likely slightly exaggerates the p values calculated from the bootstrapped null. We note that this would not affect the ordering of topics which would still lead us to include a similar set of topics in the subsequent analyses.

Conclusion

Our results thus support a model in which crime-type bias influences complex decisions through social cognitive processes. This offers a process by which biases can influence complex decisions even in the absence of evidence for strong affective experience. Thus, even after taking steps to minimize undue influence of emotional responses or moral outrage, bias-mitigation strategies in legal proceedings should also consider bias driven by social cognitive processes.

Understanding the cognitive processes underlying bias is especially important in light of the limited effectiveness of current methods for mitigating its effects in juror decisions (Ingriselli, 2014). In fact, the most effective methods for addressing racial and cultural biases generally may be tied to the use of social cognition (Paluck et al., 2021). Understanding jurors' reliance on social cognition may be important for judges and lawyers in crafting more effective means of countering bias in the justice system.

In addition to informing best practices in the legal system, we offer this study as an example of a neuroscience-based approach to elucidate a broad range of complex decision making processes in humans. Linking narrative and social cognitive processes to juror decision making may offer a template for explaining social/cultural influences in other consequential decisions (Brady et al., 2020). The approach we describe here might therefore be applied to quantifying the contributions of social cognition to biased decisions in a range of social phenomena from the spread of fake news to anti-vaccination movements.

References

Agarwal, R., Dugas, M., Ramaprasad, J., Luo, J., Li, G., and Gao, G. G. (2021). Socioeconomic privilege and political ideology are associated with racial disparity in COVID-19 vaccination. *Proc. Natl. Acad. Sci. U. S. A.* 118. doi: 10.1073/pnas.2107873118.

Allen, R. J., and Pardo, M. S. (2019). Relative plausibility and its critics. *The International Journal of Evidence & Proof* 23, 5–59.

Amodio, D., and Cikara, M. (2020). The Social Neuroscience of Prejudice. doi: 10.31234/osf.io/2sdzb.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2012). The Impact of Jury Race in Criminal Trials. *Q. J. Econ.* 127, 1017–1055.

Arkes, H. R., and Garske, J. P. (1982). *Psychological Theories of Motivation*. Brooks/Cole.

Bietti, L. M., Tilston, O., and Bangerter, A. (2019). Storytelling as Adaptive Collective Sensemaking. *Top. Cogn. Sci.* 11, 710–732.

Boccadoro, S., Cracco, E., Hudson, A. R., Bardi, L., Nijhof, A. D., Wiersema, J. R., et al. (2019). Defining the neural correlates of spontaneous theory of mind (ToM): An fMRI multi-study investigation. *Neuroimage* 203, 116193.

Bowring, A., Maumet, C., and Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Hum. Brain Mapp.* 40, 3362–3384.

Brady, W. J., Crockett, M. J., and Van Bavel, J. J. (2020). The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspect. Psychol. Sci.* 15, 978–1010.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron* 60, 930–940.

Buckholtz, J. W., and Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* 15, 655–661.

Cikara, M., and Van Bavel, J. J. (2014). The Neuroscience of Intergroup Relations: An Integrative Review. *Perspect. Psychol. Sci.* 9, 245–274.

De Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science* 305, 1254–1258.

Devine, D. J. (2012). *Jury Decision Making: The State of the Science*. NYU Press.

Devine, D. J., Krouse, P. C., Cavanaugh, C. M., and Basora, J. C. (2016). Evidentiary, extraevidentiary, and deliberation process predictors of real jury verdicts. *Law Hum. Behav.* 40, 670–682.

Dovidio, J. F., and Gaertner, S. L. (2004). "Aversive racism," in *Advances in experimental social psychology, Vol*, ed. M. P. Zanna (San Diego, CA, US: Elsevier Academic Press, x), 1–52.

Esqueda, C. W., Espinoza, R. K. E., and Culhane, S. E. (2008). The Effects of Ethnicity, SES, and Crime Status on Juror Decision Making: A Cross-Cultural Examination of Euopean American and Mexican American Mock Jurors. *Hisp. J. Behav. Sci.* 30, 181–199.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., and Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Soc. Cogn. Affect. Neurosci.* 7, 743–751.

Flanagan, F. X. (2018). Race, Gender, and Juries: Evidence from North Carolina. *The Journal of Law and Economics* 61, 189–214.

Gastwirth, J. L., and Sinclair, M. D. (2004). A re-examination of the 1966 Kalven–Zeisel study of judge–jury agreements and disagreements and their causes. *Law Probab. Risk* 3, 169–191.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., et al. (2016). Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. *J. Neurosci.* 36, 9420–9434.

Gray, K., Young, L., and Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychol. Inq.* 23, 101–124.

Greene, J., and Haidt, J. (2002). How (and where) does moral judgment work? *Trends Cogn. Sci.* 6, 517–523.

Gross, S. R., O'Brien, B., Hu, C., and Kennedy, E. H. (2014). Rate of false conviction of criminal defendants who are sentenced to death. *Proc. Natl. Acad. Sci. U. S. A.* 111, 7230–7235.

Harris, L. T., and Fiske, S. T. (2009). Social neuroscience evidence for dehumanised perception. *European Review of Social Psychology* 20, 192–231.

Hastie, R. (2019). The case for relative plausibility theory: Promising, but insufficient. *The International Journal of Evidence & Proof* 23, 134–140.

Hitchcock, R. K. (2019). Hunters and gatherers past and present: Perspectives on diversity, teaching, and information transmission. *Rev. Anthr.* 48, 5–37.

Holloway, C., and Wiener, R. L. (2018). The Role of Emotion and Motivation in Jury Decision-Making. *Criminal Juries in the 21st Century: Psychological Science and the Law*, 247.

Ingriselli, E. (2014). Mitigating jurors' racial biases: The effects of content and timing of jury instructions. *Yale LJ* 124, 1690.

Kochiyama, T., Ogihara, N., Tanabe, H. C., Kondo, O., Amano, H., Hasegawa, K., et al. (2018). Reconstructing the Neanderthal brain using computational anatomy. *Sci. Rep.* 8, 6296.

Kovel, J. (1984). *White racism: A psychohistory*. books.google.com.

Kragel, P. A., and LaBar, K. S. (2016). Decoding the Nature of Emotion in the Brain. *Trends Cogn. Sci.* 20, 444–455.

Li, R., Smith, D. V., Clithero, J. A., Venkatraman, V., Carter, R. M., and Huettel, S. A. (2017). Reason's Enemy Is Not Emotion: Engagement of Cognitive Control Networks Explains Biases in Gain/Loss Framing. *J. Neurosci.* 37, 3588–3598.

Loseke, D. R. (2007). The Study of Identity As Cultural, Institutional, Organizational, and Personal Narratives: Theoretical and Empirical Integrations. *Sociol. Q.* 48, 661–688.

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., and Meissner, C. A. (2005). Racial bias in mock juror decision-making: a meta-analytic review of defendant treatment. *Law Hum. Behav.* 29, 621–637.

Neubauer, S., Hublin, J.-J., and Gunz, P. (2018). The evolution of modern human brain shape. *Sci Adv* 4, eaao5961.

Ngo, L., Kelly, M., Coutlee, C. G., Carter, R. M., Sinnott-Armstrong, W., and Huettel, S. A. (2015). Two Distinct Moral Mechanisms for Ascribing and Denying Intentionality. *Sci. Rep.* 5, 17390.

Paluck, E. L., Porat, R., Clark, C. S., and Green, D. P. (2021). Prejudice Reduction: Progress and Challenges. *Annu. Rev. Psychol.* 72, 533–560.

Pearson, J. M., Law, J. R., Skene, J. A. G., Beskind, D. H., Vidmar, N., Ball, D. A., et al. (2018). Modelling the effects of crime type and evidence on judgments about guilt. *Nat Hum Behav* 2, 856–866.

Pennington, N., and Hastie, R. (1986). Evidence evaluation in complex decision making. *J. Pers. Soc. Psychol.* 51, 242–258.

Poldrack, R. A., Mumford, J. A., Schonberg, T., Kalar, D., Barman, B., and Yarkoni, T. (2012). Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Comput. Biol.* 8, e1002707.

Royzman, E. B., and Borislow, S. H. (2022). The puzzle of wrongless harms: Some potential concerns for dyadic morality and related accounts. *Cognition* 220, 104980.

Ruiz, J. B., and Bell, R. A. (2021). Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey. *Vaccine* 39, 1080–1086.

Shiller, R. J. (2017). Narrative Economics. *Am. Econ. Rev.* 107, 967–1004.

Smaers, J. B., and Vanier, D. R. (2019). Brain size expansion in primates and humans is explained by a selective modular expansion of the cortico-cerebellar system. *Cortex* 118, 292–305.

Smith, D., Schlaepfer, P., Major, K., Dyble, M., Page, A. E., Thompson, J., et al. (2017). Cooperation and the evolution of hunter-gatherer storytelling. *Nat. Commun.* 8, 1853.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23 Suppl 1, S208–19.

Somers, M. R. (1994). The Narrative Constitution of Identity: A Relational and Network Approach. *Theory Soc.* 23, 605–649.

Spreng, R. N., and Andrews-Hanna, J. R. (2015). The default network and social cognition. *Brain mapping: An encyclopedic reference* 1316, 165–169.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., and Sanfey, A. G. (2018). Neurobiological Mechanisms of Responding to Injustice. *J. Neurosci.* doi: 10.1523/JNEUROSCI.1242-17.2018.

Stets, J. E., and Burke, P. J. (2000). Identity Theory and Social Identity Theory. *Soc. Psychol. Q.* 63, 224–237.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., et al. (2014). Corticolimbic gating of emotion-driven punishment. *Nat. Neurosci.* 17, 1270–1275.

Unsworth, K. L., and Fielding, K. S. (2014). It's political: How the salience of one's political identity changes climate change beliefs and policy support. *Glob. Environ. Change* 27, 131–137.

Van Essen, D. C., Donahue, C. J., Coalson, T. S., Kennedy, H., Hayashi, T., and Glasser, M. F. (2019). Cerebral cortical folding, parcellation, and connectivity in humans, nonhuman primates, and mice. *Proc. Natl. Acad. Sci. U. S. A.* doi: 10.1073/pnas.1902299116.

Vidmar, N. (1997). Generic prejudice and the presumption of guilt in sex abuse trials. *Law Hum. Behav.* 21, 5–25.

Vidmar, N. (2002). Case studies of pre- and midtrial prejudice in criminal and civil litigation. *Law Hum. Behav.* 26, 73–105.

Vidmar, N. (2003). When all of us are victims: juror prejudice and terrorist trials. *Chi. - Kent L. Rev.* 78, 1143.

Vilares, I., Wesley, M. J., Ahn, W.-Y., Bonnie, R. J., Hoffman, M., Jones, O. D., et al. (2017). Predicting the knowledge-recklessness distinction in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 114, 3222–3227.

Wiener, R. L., Arnot, L., Winter, R., and Redmond, B. (2006). Generic Prejudice in the Law: Sexual Assault and Homicide. *Basic Appl. Soc. Psych.* 28, 145–155.

Willems, R. M., Nastase, S. A., and Milivojevic, B. (2020). Narratives for Neuroscience. *Trends Neurosci.* 43, 271–273.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage* 45, S173–86.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670.

Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8235–8240.

Yuan, Y., Major-Girardin, J., and Brown, S. (2018). Storytelling Is Intrinsically Mentalistic: A Functional Magnetic Resonance Imaging Study of Narrative Production across Modalities. *J. Cogn. Neurosci.* 30, 1298–1314.

Zinchenko, O. (2019). Brain responses to social punishment: a meta-analysis. *Sci. Rep.* 9, 12800.

Fig. 1. A mock-juror task tests the effects of crime scenario on bias. During the task (A), participants read a crime scenario paired with variable evidence from each of three types and rated the strength of the case against the accused and the recommended punishment severity. (B) Case-strength contributions from scenario independent of evidence. Symbols represent mean effect size (scale 0-100); error bars represent 95% credible intervals. Scenario depicted in panel A is distinguished with a triangle. Scenario effects are shown in rank order. (C) Correlation plot showing the relationship between model estimates of case strength and punishment for crime scenarios independent of evidence. This correlation (Pearson $R = 0.90$, $p < 0.001$) captures crime-type bias, which is the extent to which the seriousness of a crime increases the perceived strength of the case against the accused, independent of the evidence.

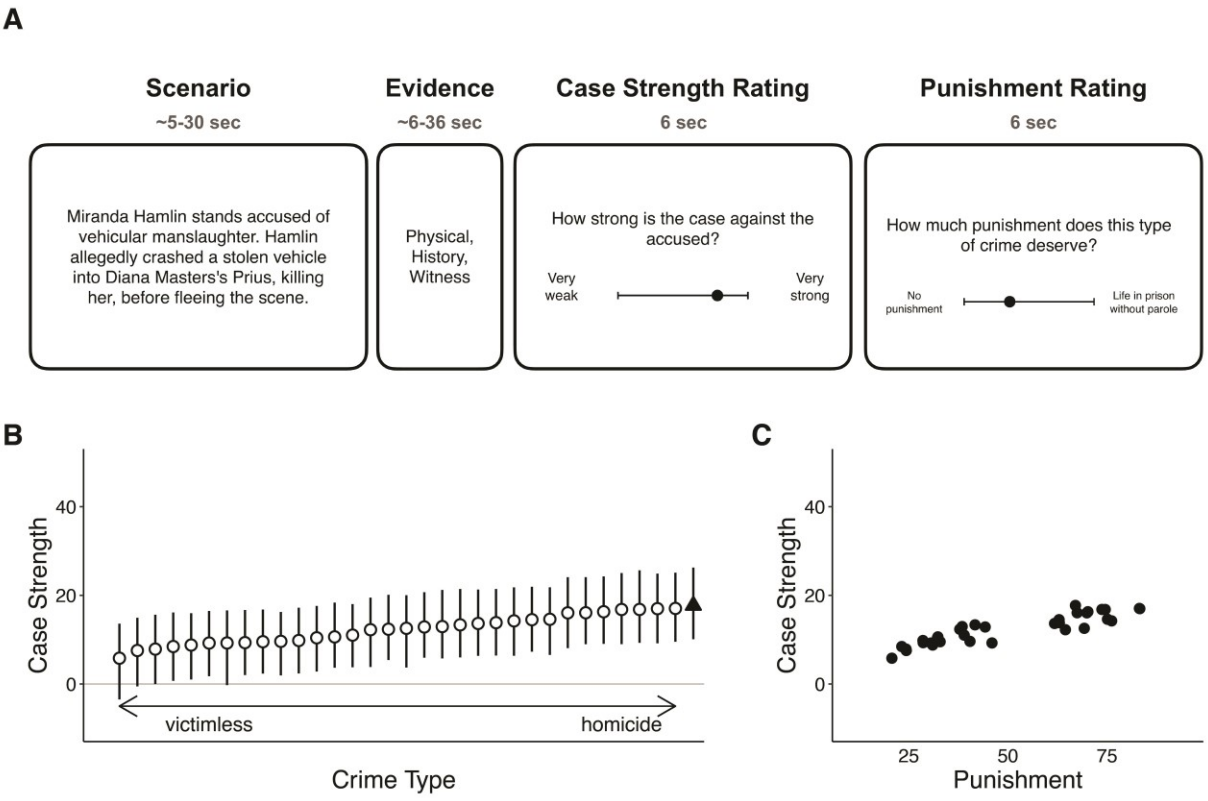

Figure 2

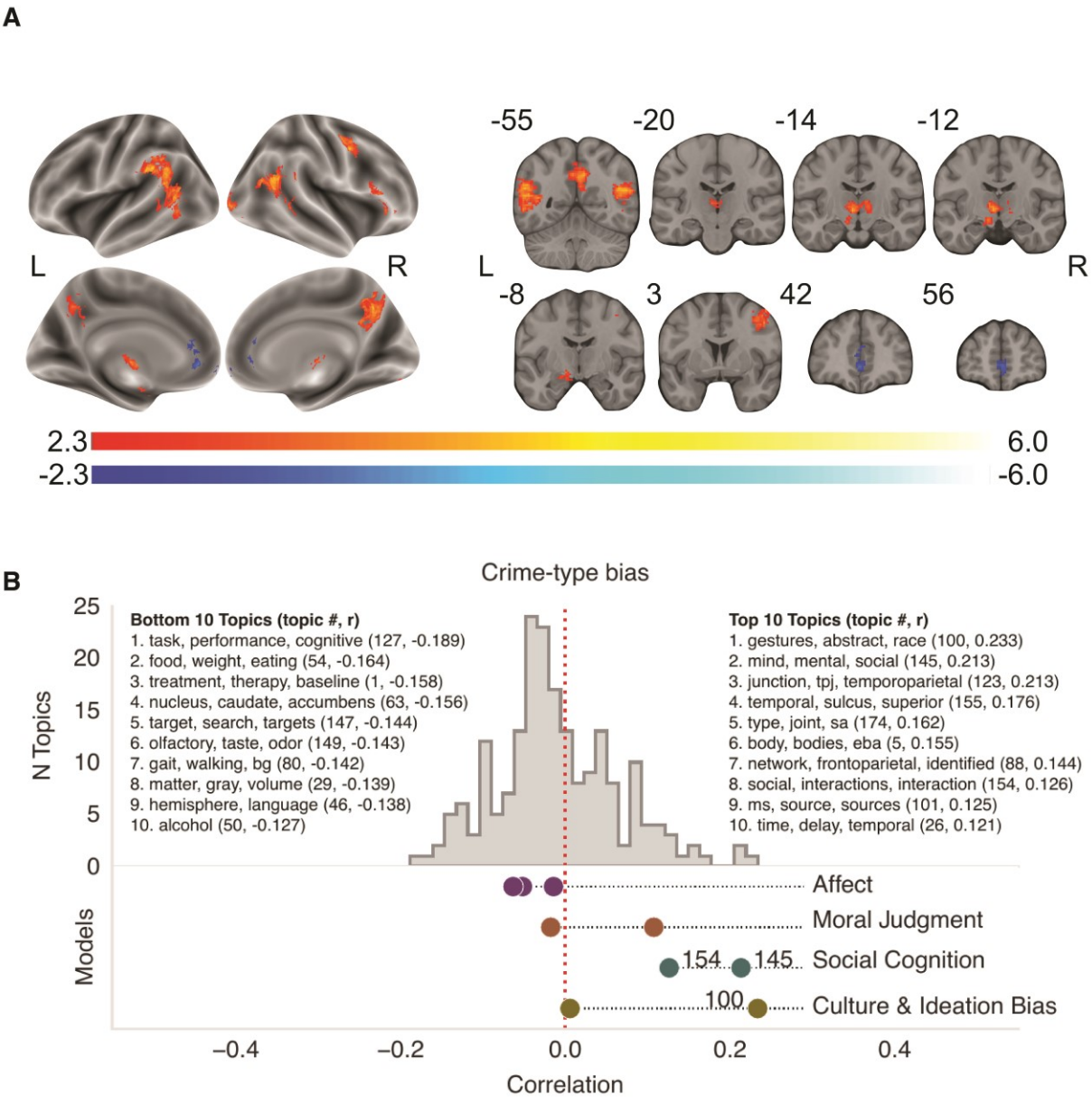

Figure 3

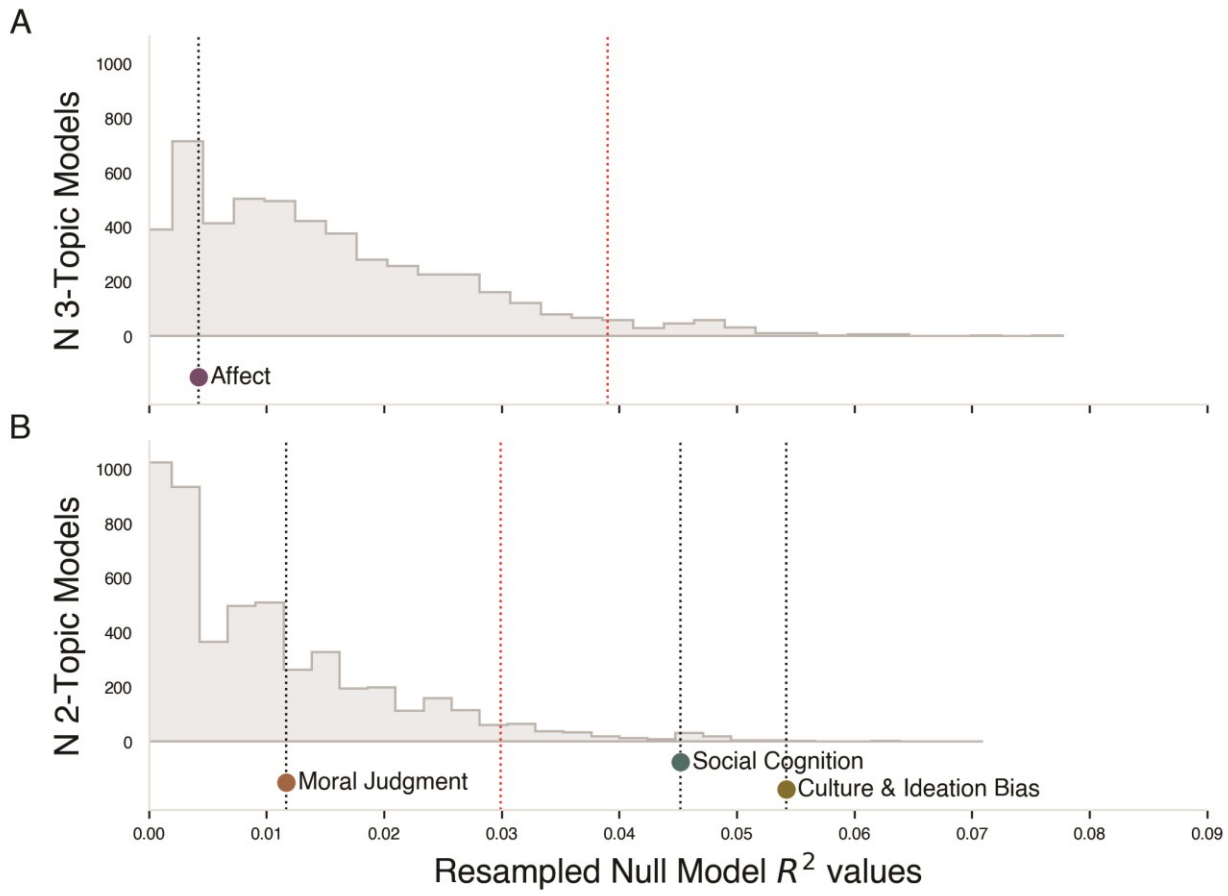

Figure 4