Humans maintain a level of cooperation among non-kin that is unrivaled in the animal kingdom. This unique feature of human culture is enabled by our ability to generate, transmit, and follow widely held agreements about morally acceptable and permissible behavior (social norms). However, the social welfare provided by cooperation crucially depends on our ability to enforce these norms by sanctioning those who violate them. Determining moral responsibility and assigning a deserved punishment are two cognitive cornerstones of norm enforcement; together, they form the foundation for modern state-administered systems of justice. Although prior work has implicated the dorsolateral prefrontal cortex (DLPFC) in social norm-based judgments, the relative contribution of this brain region to judgments of moral responsibility and punishment decision-making remains poorly understood. Here, we used repetitive transcranial magnetic stimulation (rTMS) and functional magnetic resonance imaging (fMRI) to determine the specific, causal role of DLPFC function in norm-enforcement behavior. rTMS to DLPFC significantly reduced punishment for wrongful acts without affecting blameworthiness ratings for the same acts, suggesting a neural dissociation between punishment decisions and moral responsibility judgments. We confirmed this dissociation using fMRI: DLPFC is preferentially recruited for punishment decision-making compared to blameworthiness evaluation. Finally, we employed conditional process modeling to show that DLPFC supports punishment decision-making by integrating information about culpability and harm. Together, these findings reveal a selective, causal role for lateral prefrontal cortex in punishment decision-making, and suggest a computational source for this selectivity: the ability of the DLPFC to integrate distinct information processing streams that form the basis of our punishment decisions.
The success of our species rests in large measure on our unique capacity for large-scale, stable cooperation among non-kin. Though the origin of this ability is an area of active study and debate, many attribute it to the development or elaboration of cognitive capacities that permit us to establish social norms, transmit them across generations, and detect and sanction their violation (Bendor and Swistak, 2001; Chang and Sanfey, 2013; Fehr and Fischbacher, 2004; Fehr and Rockenbach, 2004; Haushofer and Fehr, 2008; Henrich, 2006; Marlowe et al., 2011; Montague and Lohrenz, 2007; Ruff et al., 2013; Sanfey et al., 2014). Successful cooperation today is made possible by systems of justice that inflict state-authorized costs on those who would otherwise be gleeful defectors among naive or resigned cooperators. Indeed, regardless of the specific phylogeny of human ultra-sociality, the continued stability of modern human societies hinges on our ability to enforce widely shared sentiments about appropriate behavior (Buckholtz and Marois, 2012; Fehr and Fischbacher, 2004; Marlowe et al., 2011).

Given the importance of norms for the development of modern human culture, some have suggested that human brains are especially well equipped to make norm-based judgments (Buckholtz and Marois, 2012; Crockett, 2013; Fehr and Camerer, 2007; Sanfey et al., 2014). Over the last decade, functional imaging and brain stimulation work implicate one region in particular - the dorsolateral prefrontal cortex (DLPFC; specifically, the anterior aspect of brodmann area 46 sometimes referred to as rostrolateral PFC) - as being crucial for norm-enforcement. These studies show evidence of DLPFC engagement across a variety of tasks indexing moral decision-making (Cushman et al., 2012; Prehn et al., 2007; Tassy et al., 2012), second-party punishment (Knoch et al., 2010; Knoch et al., 2006; Sanfey, 2003), third-party punishment (Buckholtz et al., 2008; Schleim et al., 2011; Strobel et al., 2011), and norm compliance (Baumgartner et al., 2011; Chang et al., 2011; Ruff et al., 2013; Strobel et al., 2011). Some have suggested that the involvement of DLPFC across these tasks reflects cognitive control (Haushofer and Fehr, 2008; Knoch et al., 2006; Tassy et al., 2012), while others have posited the DLPFC is necessary to assign causal responsibility to agents during norm-based judgments (Fugelsang and Dunbar, 2005; Roser et al., 2005; Satpute et al., 2005).

However, the complex nature of the norm-enforcement construct itself makes it challenging to pin down the precise role of DLPFC. Norm enforcement is not a single, unitary cognitive process, but rather comprises a range of distinct subcomponent processes. These include evaluating an agent's action with respect to shared codes for acceptable conduct (moral permissibility); assessing the agent's role in causing that act (causal responsibility); determining the agent's mental state during the act - especially his intentions (moral responsibility or “blameworthiness”); appraising the outcome of the act, particularly whether and how much it harmed other people (harm assessment); and finally, arriving at an appropriate sanction for the act (punishment) (Buckholtz and Marois, 2012; Cushman, 2008; Gray et al., 2012). The challenge inherent in parsing this construct experimentally has made it difficult to selectively map DLPFC to a specific cognitive subdomain within the larger norm-enforcement construct.

Interestingly, the particular area of DLPFC engaged during norm-enforcement has also been consistently observed in a range of non-social paradigms. Across the cognitive tasks in...
which activation in this region has been reported - such as working memory, analogical reasoning, and rule-based decision-making - the unifying feature appears to be a requirement to integrate representations from multiple subtasks in order to select responses that are adaptively matched to task goals (Bunge et al., 2005b; Christoff et al., 2001; De Pisapia and Braver, 2008; De Pisapia et al., 2012; De Pisapia et al., 2007; Duncan, 2010; Hampshire et al., 2011). Extending this conceptualization of DLPFC to the domain of norm-enforcement, we have recently offered an integration-and-selection hypothesis for the role of DLPFC in norm based judgments (Buckholtz and Marois, 2012). According to this hypothesis, DLPFC is responsible for integrating output representations from decision-relevant subtasks during norm-enforcement; this integrated signal is then used to bias response selection toward the most contextually appropriate action. For example, retributive punishment in legal contexts requires the integration of at least two main streams of information to arrive at a just sanction: 1) the severity of the criminal offense (i.e., the harm it caused); and 2) the blameworthiness of the offender (i.e., his/her moral culpability, as a function of his state of mind at the time of the offense) (Darley, 2009; LaFave, 2010). Punishment judgments are therefore the final output of an integration process that jointly considers information about the harm a defendant caused and information about his/her perceived blameworthiness for having caused it. DLPFC recruitment during decisions to punish culpable agents for criminal violations (Buckholtz et al., 2008; Treadway et al., 2014), its reported involvement in second-party economic norm-enforcement paradigms (Knoch et al., 2006; Sanfey, 2003), and its consistency across multiple classes of norms (e.g. fairness and distributive norms, moral norms, and laws) (Cushman et al., 2012; Schleim et al., 2011; Strobel et al., 2011; Tassy et al., 2012) have led us to propose that DLPFC acts as a superordinate processing node that receives and integrates context-dependent “biasing” inputs during norm-enforcement decisions. Based on prior work, we speculate that these inputs arise from medial corticolimbic circuitry and the temporo-parietal junction (TPJ), encoding harm and blameworthiness signals respectively (Buckholtz et al., 2008; Li et al., 2009; Yoder and Decety, 2014; Treadway et al., 2014). According to this model, the output of this putative DLPFC integration process, reflecting an interaction between harm and culpability (i.e. the degree of moral blameworthiness attending to an agent’s actions), biases selection from among an array of context-specific punishment response options (Buckholtz and Marois, 2012).

This hypothesis makes several testable predictions. First, DLPFC should be particularly sensitive to decisions that require joint consideration of moral responsibility and harm severity, compared to decisions that rely only on one of them. Specifically, we predict that the involvement of DLPFC should be more evident when participants are asked to determine an appropriate punishment for a norm violation as compared to when they are only instructed to rate an agent’s moral responsibility (blameworthiness) for that norm violation. This prediction is grounded in the fact that punishment decisions require representational integration from (at least) two processes (mental state evaluation and harm assessment), while blameworthiness assessments primarily hinge on mental state representations. Second, we predict that the involvement of DLPFC should be more pronounced for punishment decisions about blameworthy agents, compared to agents for whom responsibility has been mitigated by an extenuating circumstance. This prediction is grounded in the idea that...
responsibility judgments precede and constrain harm-based judgment (Buckholtz et al., 2008; Buckholtz and Marois, 2012; Treadway et al., 2014, but see Leslie et al., 2006). In other words, once information that reduces or eliminates the culpability of an agent is presented, information about harm is largely irrelevant to punishment, alleviating the need to integrate this information with intent representations in order to make an appropriate judgment.

In the present study, we combine brain stimulation and neuroimaging approaches to 1) identify a selective role for DLPFC in blameworthiness vs. punishment, and 2) test the integrative model of DLPFC function in norm-enforcement. To that end, we exploited the fact that punishment and blameworthiness judgments have distinct information processing requirements; the former requires a decision-maker to integrate representational outputs from mental state evaluation and harm severity assessment, while the latter is simply the product of mental state evaluation. First, we used a between-groups, sham-controlled repetitive transcranial magnetic stimulation (rTMS) paradigm targeting DLPFC in a sample of 66 healthy volunteers. In two separate sessions, participants were asked to make punishment and blameworthiness decisions for each of a series of hypothetical textual scenarios in which a protagonist ('John') commits a crime. These scenarios varied both in the harm caused by the criminal act and the protagonist's culpability for having caused it. Harms ranged from simple theft to murder, while culpability varied according to the protagonist's mental state: in some trials “John” could be held fully responsible for his actions (R trials); in other trials, duress, psychosis or other mitigating circumstances resulted in Diminished Responsibility for his otherwise criminal behavior (DR trials). The punishment task entailed deciding how much punishment John deserved for his actions, whereas the blameworthiness task asked how morally responsible John was for his actions. We predicted that DLPFC TMS would affect punishment – but not blameworthiness – decisions, and would do so by interfering with the appropriate integration of harm and culpability signals during decision-making. We then used fMRI in a separate cohort of subjects to provide multi-modal convergent evidence for the selective involvement of DLPFC in punishment decision-making. If the hypothesized integration function of the DLPFC in punishment decision-making were correct, we would expect greater DLPFC activity when participants make punishment decisions compared to blameworthiness judgments.

**Results**

**Culpability and Harm Severity Increase Punishment and Blameworthiness Assignment**

We first examined the impact of culpability and harm severity on blameworthiness and punishment ratings within the entire sample (i.e. collapsed across sham and active conditions) by performing separate Repeated-Measures Analysis of Variance (RM-ANOVA) for each judgment type (Punishment & Blameworthiness), with Culpability (R vs. DR) and Harm Severity as within-subject factors (See Supplemental data and Supplementary Table 1 for response time data). Harm Severity was dummy-coded as an ordinal variable according to scenario offense: 1 = Property Crime (theft and property damage), 2 = Physical Harm.
(simple assault), 3 = Severe Physical Harm (maiming, rapes), 4 = Murder (murder, including combined rape and murder).

We found a significant effect of Culpability on both Punishment and Blameworthiness ratings (Figure 1A-B). Across all participants, both Punishment and Blameworthiness ratings were higher for Responsibility trials compared to Diminished Responsibility trials (Punishment: $F_{1,59} = 1508.62, p < 0.001$; Blameworthiness: $F_{1,59} < 120.99, p = 0.001$; test of within-subject effects from RM-ANOVA; Supplementary Table 1). The main effect of Harm Severity was also significant for both Punishment ($F_{3,177} = 221.76, p < 0.001$) and Blameworthiness ($F_{3,179} = 22.24, p < 0.001$), with higher Harm Severity associated with higher ratings for both. The effect of Culpability and Harm Severity on Punishment and Blameworthiness ratings remained significant when controlling for participant gender and scenario set (see Methods; $p < 0.001$ for punishment ratings, $p < 0.05$ for blameworthiness ratings).

Greater integration of Culpability and Harm During Punishment

The integration-and-selection hypothesis implies that information about Culpability and Harm interact to affect Punishment. Supporting this notion, we found a significant Culpability-by-Harm Severity interaction for Punishment $F_{3,77} = 100.93$. This interaction effect appears to be driven by the steeper increase in Punishment ratings per increase in Harm Severity for R trials compared to DR trials. (Figure 1B,D). We did observe a statistically significant Culpability-by-Harm Severity interaction for Blameworthiness ratings as well ($F_{3,77} = 4.74, p = 0.003$), suggesting that these two factors do interact to affect Blameworthiness assessment (see Discussion). However, the integration-and-selection hypothesis not only implies that information about Culpability and Harm interacts to affect norm-enforcement, it also predicts that this interaction will be stronger for Punishment than for Blameworthiness judgments. Consistent with this hypothesis, the Culpability-by-Harm Severity interaction was significantly stronger for Punishment decisions compared to Blameworthiness judgments (Judgment Type-by-Responsibility-by-Harm Severity 3-way interaction: $F_{3,177} = 22.04, p < 0.001$). Similarly, effect sizes for the Culpability-by-Harm Severity interaction were larger for Punishment decision-making (partial $\eta^2 = 0.63$) than Blameworthiness evaluation (partial $\eta^2 = 0.07$). Visual inspection of the interaction plots corroborate these statistical analyses: for Punishment, there was a markedly steeper increase in the amount of assigned punishment per unit increase in Harm severity, but only for trials in which the protagonist was fully responsible (Fig 1D). By contrast, Blameworthiness judgments were better characterized by a step function, with relatively small differences between Harm levels within R and DR trials, accompanied by a very large difference in Blameworthiness ratings between R and DR trials (Fig 1C). Taken as a whole, these results are consistent with our supposition that integration demands are significantly higher for Punishment decisions as compared to Blameworthiness judgments.

Transient Disruption of DLPFC Selectively Alters Punishment Decisions

DLPFC function was focally and transiently disrupted with rTMS. We applied 30 minutes of 1Hz stimulation (Active group) or sham stimulation (Sham group) to left or right hemisphere DLPFC (see Fig. 2 and Methods). To determine the disruptive effect of
stimulation, we used a series of linear mixed effect models. Subject was treated as a random effect and rTMS stimulation condition, stimulation hemisphere, Culpability, and Harm Severity and rating were modeled as fixed effect predictors. The first set of models examined the simple effect of rTMS stimulation condition on Blameworthiness and on Punishment ratings (i.e. Blameworthiness and Punishment trials modeled separately), after accounting for variance due to Culpability and Harm Severity. The second set of models examined interactions between rTMS stimulation condition and stimulation hemisphere on Blameworthiness and on Punishment, after accounting for variance due to Culpability and Harm Severity. The third set of models examined rTMS-x-Culpability and rTMS-x-Harm Severity interactions on Blameworthiness and on Punishment. The final model tested rTMS stimulation-x-Judgment Type interactions. Parameter estimates were obtained via restricted maximum likelihood estimation. Gender and scenario set were included as covariates in all models. We did not detect an effect of rTMS stimulation on response times (see Supplementary Table 1).

For Blameworthiness ratings, the main effect of rTMS condition was not significant ($F_{1,55} = 0.031, p = 0.86$), nor was the Culpability by rTMS Condition interaction ($F_{1,1609} = 2.57, p = 0.11$). In addition, we did not observe a significant effect of Hemisphere ($F_{1,54} = 0.02, p = 0.88$), and the two-way (Hemisphere by rTMS Condition) and three-way (Hemisphere by rTMS Condition by Culpability) interactions were also not significant ($p's > 0.25$) (Fig 3A).

By contrast, we found a significant main effect of rTMS Condition ($F_{1,55} = 4.37, p = 0.04$), and a significant Culpability by rTMS Condition interaction ($F_{1,1609} = 9.94, p = 0.002$) for Punishment decisions. We did not observe a significant main effect for Hemisphere, nor did we observe significant two-way (Hemisphere by rTMS Condition) or three-way (Hemisphere by rTMS Condition by Culpability) interactions (all $p$-values > 0.05).

Descriptively (Fig 3B), punishment ratings were lower in the Active compared to the Sham rTMS Condition for Responsibility trials (5.74±0.15 vs. 6.33±0.15) but did not differ by rTMS Condition for Diminished Responsibility trials (0.94±0.13 vs. 0.98± 0.15, Active vs. Sham, respectively). Post-hoc comparisons confirmed that the effect of rTMS was significant for Responsibility trials ($F_{1,54} = 7.78, p = 0.007$), while no such effect was observed for Diminished Responsibility trials ($F_{1,54} = 0.05, p = 0.83$). These data indicate that DLPFC rTMS significantly reduced punishment for culpable criminal acts, and that the magnitude of this effect did not differ as a function of which hemisphere was stimulated.

Thus, supporting our hypothesis, DLPFC rTMS selectively affected Punishment decisions about culpable agents, but not judgments of Blameworthiness. To further test this selectivity, we submitted the Punishment and Blameworthiness values for each group to a formal interaction test. This analysis revealed a significant rTMS Condition (Active vs. Sham) by Judgment type (Punishment vs. Blameworthiness) interaction ($F_{1,3294} = 4.82, p = 0.028$), such that the effect of rTMS was significantly larger for Punishment judgments compared to Blameworthiness judgments.

Taken together, these data confirm a causal role for DLPFC in third-party norm-enforcement that is selective for punishment decision-making, as DLPFC rTMS did not affect blameworthiness judgments. Specifically, disrupting DLPFC function lowered the amount of
punishment assigned for Responsibility scenarios. To further unpack this effect, we examined the effect of rTMS on mean punishment ratings at each level of Harm Severity separately for R and DR scenarios (Fig 3C-D). Consistent with the analyses reported above, rTMS did not significantly alter punishment at any level of Harm Severity for DR scenarios (i.e. when the agent’s culpability was reduced by extenuating circumstances; all p-values >0.3 for all harm levels; Fig 3D). For fully culpable agents (R scenarios), we found that the effect of rTMS was stronger for low-harm compared to high-harm crimes (Property Crime: p = 0.004; Assault: p = 0.05; Maim and Rape: p = 0.20; Murder: p = 0.47). As expected, rTMS did not modulate blameworthiness ratings at any level of Harm Severity for either R or DR scenarios (all p-values > 0.37; Fig 3C). These data indicate that disrupting DLPFC function lowers punishment for culpable agents, but only when their norm violations result in low-moderate harm.

DLPFC TMS Disrupts The Integration of Culpability and Harm During Punishment but not Blameworthiness Decisions

The above results suggest that DLPFC rTMS may impair the utilization of mental state information during punishment decision-making in a manner that is harm-sensitive. This accords well with our prediction that DLPFC rTMS would disrupt the joint consideration of Culpability and Harm during punishment decision-making (Buckholtz and Marois, 2012). To further unpack the mechanisms through which DLPFC disruption affects punishment decisions, we ran a series of regression models to estimate the relative influence of mental state and harm information on punishment decisions for each subject, and compared these estimates between rTMS groups.

Punishment Models: Culpability and Harm—Across all participants and all trials, both Culpability and Harm Severity were significant, unique predictors of punishment amount (Model 1, Culpability: $\beta_{\text{Culpability}} = -0.77, p < 0.001$; Model 2, Culpability, Harm Severity: $\beta_{\text{Culpability}} = -0.77, p < 0.001, \beta_{\text{Harm}} = 0.32, p < 0.001$; Model 1 $R^2 = 0.6$, Model 2 $R^2 = 0.71, R^2$ change = 0.1, p < 0.001). We then calculated punishment-based Culpability and Harm beta-weights for each subject. Across participants, Culpability and Harm beta-weights were negatively correlated (Pearson $r = -0.64$, p <0.001), suggesting that subjects differ in the relative weight they accord to culpability and harm in their punishment decisions: that is, for a given individual, when the influence of culpability on punishment is high, the influence of harm tends to be low (and vice versa) (Fig 4A).

Punishment Models: TMS Effects—We used a multivariate general linear model analysis to compare participants’ $\beta_{\text{Culpability}}$ and $\beta_{\text{Harm}}$ values across rTMS groups (hemisphere and rTMS Condition were included as fixed factors, and sex was included as a covariate). Notably, $\beta_{\text{Culpability}}$ values were significantly lower in the active, compared to sham rTMS groups ($F_{1,55} = 4.13, p = 0.047$ for main effect of rTMS; p’s > 0.15 for main effect of Hemisphere and for rTMS-by-Hemisphere interaction). By contrast, $\beta_{\text{Harm}}$ values were significantly higher in participants exposed to DLPFC rTMS ($F_{1,55} = 6.69, p = 0.01$ for main effect of rTMS; p > 0.7 for main effect of hemisphere and for TMS-by-hemisphere interaction) (Fig. 4B). This suggests that disrupting DLPFC function attenuates the impact of information about offender culpability while simultaneously potentiating the influence of
information about harm on punishment. As an additional test of the hypothesis that DLPFC supports the integration of culpability and harm signals, we constructed the multiplicative interaction term $\beta_{\text{Culpability}} \times \beta_{\text{Harm}}$. This term was significantly different between rTMS groups ($p < 0.05$), providing further support for the notion that DLPFC performs an integration-and-selection function during third-party norm-enforcement.

**Punishment Models: Mediation Analyses**—If DLPFC rTMS affects punishment decision-making by interfering with the integration of signals for culpability and harm, then we would expect the impact of DLPFC rTMS on punishment to be mediated by these signals (Fig 5). We tested this hypothesis using a mediation analysis with rTMS condition as a predictor of punishment scores in Culpability trials, $\beta_{\text{Culpability}}$ and $\beta_{\text{Harm}}$ as mediators, and gender and scenario set as nuisance covariates. The total effect model was significant ($F_{1,56} = 3.74, p = 0.02$), as was the total effect of rTMS on punishment ($\beta_{\text{total}} = 0.57; 0.15-0.98, 95\% \text{ C.I.}$). Decomposing the total effect, we found that the direct effect of rTMS on punishment in this model was not significant ($\beta_{\text{direct}} = 0.16; -0.14-0.46, 95\% \text{ C.I.}$); however we did observe indirect effects through the two mediators ($\beta_{\text{Culpability}} = 0.17; 0.003-0.39, 95\% \text{ C.I.}$; $\beta_{\text{Harm}} = 0.24; 0.02-0.49, 95\% \text{ C.I.}$). These data confirm our hypothesis that TMS modulates punishment by affecting the way that participants use information about culpability and harm in selecting appropriate third-party sanctions for norm violations.

**Blameworthiness Models: $\beta$-weight and mediation analyses**—While the regression models and mediation analysis confirmed that DLPFC rTMS disrupts the integration of Harm and Culpability during punishment decision-making, no such disruptive effect of rTMS on signal integration was expected for Blameworthiness judgments. Across all participants and all trials, Culpability and Harm Severity were significant predictors of blameworthiness (Model 1, Responsibility: $\beta_{\text{Responsibility}} = -0.71, p < 0.001$; Model 2, Responsibility, Harm Severity: $\beta_{\text{Responsibility}} = 0.71, p < 0.001$, $\beta_{\text{Harm}} = 0.13, p < 0.001$; Model 1 $R^2 = 0.50$, Model 2 $R^2 = 0.51$, $R^2$ change = 0.018, $p < 0.001$). However, in contrast to decisions about punishment, blameworthiness $\beta_{\text{Responsibility}}$ values did not differ between TMS conditions (Active vs. Sham: $p = 0.99$; Hemisphere: $p = 0.60$; TMS Condition-By-Hemisphere interaction: $p = 0.54$), nor did $\beta_{\text{Harm}}$ values (Active vs. Sham: $p = 0.50$; Hemisphere: $p = 0.81$; TMS Condition-By-Hemisphere interaction: $p = 0.70$) or any of the interaction beta-weights described above ($p > 0.5$, all comparisons). The absence of rTMS effects obviated the need to perform mediation analyses for blameworthiness trials. Thus, while harm and culpability each account for unique variance in blameworthiness ratings, the influence of these factors on blameworthiness judgments is not affected by DLPFC rTMS.

**DLPFC Activity is Selective for Punishment Decision-Making**

The TMS data above indicate that transient disruption of DLPFC function affects norm-enforcement behavior. Further, the exclusive effect of DLPFC on punishment decision-making (compared to blameworthiness judgments) suggests a relatively selective mapping between DLPFC function and one specific cognitive component of norm-enforcement behavior, namely punishment decision-making. As a convergent test of this apparent selectivity, we compared DLPFC activity with fMRI in 10 participants who were asked to make punishment and blameworthiness judgments for the same set of scenarios used in the
TMS experiment. Using blood oxygen level-dependent (BOLD) signal extracted from our left and right DLPFC TMS stimulation target sites (Fig. 6A, see Methods), we tested for DLPFC activation differences between punishment decisions and blameworthiness judgments. Consistent with prior work (Buckholtz et al., 2008), we found a significant main effect of Responsibility in right DLPFC ($t_{9} = 2.41, p = 0.04$), such that the BOLD signal in this brain region was higher during R than during DR trials (Fig. 6B). This relationship was also observed in left DLPFC, albeit marginally ($t_{9} = 2.12, p = 0.06$). Importantly, no such relationship was observed for blameworthiness judgments in either the right (Fig. 6B) or left DLPFC ($p's > 0.4$). Moreover, we found a significant judgment type difference ($t_{9} = -2.23, p = 0.05$), with right DLPFC more active during punishment decisions compared to blameworthiness judgments. This difference was not observed in left DLPFC ($t_{9} = -0.18, p = 0.86$). Taken together with the behavioral results (see Fig. 1), these findings suggest that while Punishment decisions and Blameworthiness judgments are both sensitive to culpability differences, the use of culpability information by DLPFC is selective for Punishment.

Discussion

The principal finding of the present study is that inhibitory transcranial magnetic stimulation of the DLPFC reduces the punishment of culpable agents without affecting judgments of their blameworthiness. Norm-enforcement involves assigning blameworthiness to a norm violator based on an evaluation of causal responsibility and mental state, assessing the outcome of the norm violation (i.e. the magnitude of harm), and combining these calculations to arrive at an appropriate sanction (Buckholtz and Marois, 2012). The current rTMS experiment confirms that assessing blameworthiness and assigning punishment are cognitively distinct processes, with DLPFC involvement selective for the latter. fMRI provides convergent evidence for this selectivity, with (right) DLPFC activity sensitive to culpability differences during decisions about punishment but not about blameworthiness. We postulate that blameworthiness judgments are a temporally antecedent (and perhaps prerequisite) process, the output of which (i.e. culpability estimates) is used to calibrate the impact of harm severity on punishment magnitude selection. Together, these data demonstrate a selective, causal role for DLPFC in norm-enforcement.

DLPFC is a cortical area that has undergone significant expansion in size, specialization, and connectivity through hominoid evolution, with striking differences evident between humans and other apes (Sakai et al., 2011; Semendeferi et al., 2011a; Semendeferi et al., 2011b). While it would seem unlikely that DLPFC (or a portion thereof) is specialized for norm-enforcement broadly, or punishment specifically, domain-general aspects of cognition that were enabled or enhanced by DLPFC expansion are likely necessary to support this process. The DLPFC sends and receives projections from other multimodal association areas, motor cortex and subcortical zones, making it well suited to coordinate a variety of complex processes (Duncan, 2010; Mesulam, 1998; Miller and Cohen, 2001). A key property of DLPFC microcircuits is their ability to maintain stable representations over time (Dehaene and Changeux, 1995; Goldman-Rakic, 1990), underlying the role of this region in working memory (Braver et al., 1997; Goldman-Rakic, 1990; Nee et al., 2013). This property, in part, enables DLPFC to coordinate among available external sensory inputs, internal states, and
response options in the service of promoting adaptive behavior (Duncan, 2010; Fuster, 1993; Mesulam, 1998; Miller and Cohen, 2001; Passingham and Sakai, 2004). The massively integrative nature of information processing within DLPFC likely explains why it appears to be fundamental to many aspects of higher-order cognition and decision-making (Barbas and Zikopoulos, 2007; Bunge et al., 2005a; Fuster, 1993). More recent work suggests that the region of PFC that we have targeted in the present study is specifically recruited when adaptive performance requires that abstract representations maintained in distinct information processing streams be synthesized and integrated into ongoing behavior (Bunge et al., 2005b; Christoff et al., 2001; De Pisapia and Braver, 2008; De Pisapia et al., 2007; Koechlin et al., 1999; Koechlin and Summerfield, 2007; Nee et al., 2013). While many of these studies have characterized the integrative functions subserved by DLPFC in the context of working memory and relatively simple cognitive tasks, we believe that this same DLPFC-dependent, domain-general integration process also enables significantly more complex forms of behavior, such as norm-enforcement.

According to our integration-and-selection model (Buckholtz and Marois, 2012), DLPFC combines information about harm and culpability with context-specific punishment rules (e.g. norm-specific punishment scales and culture-specific mitigating circumstances). The current findings offer support for this model in several key ways. First, we show that harm and culpability interact to determine punishment magnitude. Specifically, our suggestion that norm-based punishment requires a synthesis of harm and culpability signals is supported by the finding of a significant harm-by-responsibility interaction for punishment values, as well as a significant negative correlation between beta-weights for harm and culpability predictors. Importantly, DLPFC rTMS interferes with this synthesis. First, rTMS attenuated the influence of culpability information while simultaneously increasing the influence of harm severity signals on punishment. Moreover, disrupting DLPFC affected the interaction between harm and culpability signals. Finally, mediation analysis confirms that the impact of DLPFC rTMS affect punishment by altering the effect of harm and culpability information on punishment magnitude. As a whole, these findings are consistent with the notion that DLPFC supports norm-enforcement by synthesizing decision-relevant streams of information in order to bias selection from among competing response options.

Several outstanding issues merit further consideration. First, we observed that DLPFC rTMS decreases mean punishment scores. On its face, this is similar to Knoch and colleagues' finding of diminished second-party punishment in the ultimatum game following DLPFC rTMS (Knoch et al., 2006). While those authors attribute decreased second-party punishment to a rTMS-induced cognitive control impairment – namely, reduced inhibition of the prepotent response to receive monetary gains - this explanation is difficult to reconcile with the present data. Here, one might expect reduced inhibitory control to manifest primarily in the DR condition, when mitigating information induces a requirement to inhibit the prepotent response to punish those that have harmed others. However, significant effects of rTMS on punishment are only observed for R scenarios in the current dataset, which would not be predicted by an inhibitory control account of DLPFC function during norm enforcement. Nevertheless, while our study emphasizes the disruptive effects of DLPFC rTMS on information integration and response selection during punishment, the current data do not rule out the involvement of other related DLPFC-dependent processes (e.g. inhibitory
control) in norm enforcement behavior (Haushofer and Fehr, 2008); indeed, it is plausible to suggest that both integration and inhibitory control operate in tandem during norm-enforcement. A more direct test of the relationship between integration-and-selection and inhibitory control during punishment awaits future work.

Second, while our findings directly implicate rDLPFC in punishment decisions, a finding that is consistent with our prior correlational studies (Buckholtz et al., 2008; Treadway et al., 2014), they do not speak to whether it is the only brain region involved in this process. rTMS is, technically speaking, a focal methodology, and our fMRI experimental design was a priori designed to specifically assess the role of DLPFC in punishment and blameworthiness judgments. Thus, it is possible that other brain regions are differentially involved in these two judgments. Consistent with this idea, prior work has associated mental state inference, a key component of moral judgment and presumably blameworthiness, with temporoparietal junction activity rather than DLPFC activity (Decety and Cowell, 2014; Yoder and Decety, 2014; Young et al., 2010; Young et al., 2007; Treadway et al., 2014). On the whole, norm-enforcement behavior is likely facilitated by complex functional interactions between multiple brain regions subserving different cognitive computations (Buckholtz & Marois, 2012; Treadway et al., 2014; Buckholtz and Meyer-Lindenberg 2012; Buckholtz 2015).

Third, DLPFC rTMS appears to reduce punishment by simultaneously diminishing the influence of information about culpability and enhancing the influence of information about harm severity. At first glance, one might expect that boosting the impact of harm signals would increase punishment. However, the punishment-reducing effect of TMS is only observed for low-moderate harms. For acts that result in such harms, considering the outcome may result in lower punishment than considering the malicious intent that produced that outcome. In other words, reliance on harm signals would produce a lower punishment because the actual harm that occurred is of low magnitude, while relying on culpability assessment could result in a higher punishment because the norm enforcer is punishing based on the agent’s intentions (or perhaps, on the outcome that they believe the agent desired) rather than the actual low-harm outcome. Future modeling work on punishment decision-making will help better elucidate the precise nature of the computations that lead to punishment, and the role that specific circuits play in representing these computations.

Finally, we note that the mean effect of DLPFC rTMS is relatively modest. This may be due to the fact that we used a group-based coordinate that we targeted on MNI normalized brains. Functional localization of subject-specific DLPFC foci may prove a more powerful approach to stimulation-based behavioral modulation(Saxe et al., 2006). This technique, in combination with alternative brain stimulation methods that permit both up-regulation and down-regulation of cortical function (e.g. transcranial direct current stimulation) offers a particularly compelling approach to parsing the neural circuitry involved in norm-enforcement behavior(Ruff et al., 2013).

Using non-invasive cortical stimulation and fMRI, we outline here a domain-general, DLPFC-dependent cognitive mechanism – integration-and-selection – underlying third-party punishment decisions for social norm violations (crimes). The current data suggests a possible neuroanatomical parsing of norm-enforcement, with DLPFC function selectively
mapping to one component of this construct (assigning deserved punishment) but not another (assessing moral responsibility). The dissociation between punishment and blameworthiness observed here accords well with previous studies on the second-party punishment of distributional (fairness) norms, in which DLPFC disruption appeared to selectively affect the punishment of intentional norm violations (Knoch et al., 2006; Sanfey, 2003) while leaving intact an evaluation of the fairness of an agent’s behavior and the presence of a norm violation (see also Ruff et al., 2013 for a similar dissociation for norm compliance) (Ruff et al., 2013). The fact that rTMS did not disrupt blameworthiness judgments in the present experiment is all the more remarkable given that that this judgment utilized a response scale that was identical to the one used for the punishment decision, differing only in the type of decision (degree of moral responsibility for an act versus appropriate punishment for that act). This finding in turn suggests that high-level evaluative and reasoning processes that are crucial for norm-enforcement (assessment of moral responsibility) may take place with minimal involvement of the DLPFC (at least the DLPFC area targeted in the present study). Indeed, it may be that the DLPFC supports norm enforcement not by instantiating any one particular cognitive process, but rather by integrating the outputs of a variety of norm-relevant cognitive processes.

The current study provides a suggestive window into the cognitive mechanisms that underlie paradigmatic decisions in the criminal justice system. Modern institutions of justice depend on the ability of disinterested third-parties – typically jurors and judges – to integrate information about the actions and mental states of others in order to decide whether to punish and, if so, how much (Bendor and Swistak, 2001; Boyd et al., 2010; Crockett et al., 2014; Henrich, 2006; Marlowe et al., 2011). Thousands of jurors and judges weigh the fates of criminal defendants every day, a process that enables the large-scale cooperation and widespread peace that we all enjoy. However, this process is no less remarkable for being so commonplace. While *Homo sapiens* is not the only primate species to punish in retaliation for direct harms (second-party punishment) (Jensen et al., 2007a, b; Proctor et al., 2013), humans alone among all animals appear willing to bear personal costs in order to sanction those who have harmed others (Riedl et al., 2012). The adoption of this norm enforcement strategy by our species is thought to have played a crucial role in the evolutionary stability of human cooperation (Bendor and Swistak, 2001). This study nominates DLPFC, a region that is uniquely suited for representational integration, as a core neural substrate for this capacity.

**Experimental Procedures**

**rTMS Study**

**Participants**—Sixty-six healthy volunteers (aged 18-30; 32 males) were recruited from the Vanderbilt University community through the Department of Psychology’s Study Pool website. All participants provided written informed consent, and all study procedures were approved by the Vanderbilt University Institutional Review Board. Exclusion criteria included current treatment with psychoactive medication, history of major psychiatric illness, diagnosed neurological problems, less than high-school level of education, non-native English language speaker, pregnancy, or left-handedness. Additionally, participants who reported either being the victim of a physical or sexual crime, or being a witness to such
crime in the last two years were excluded. Participants who have experienced such a crime more than two years ago but expressed recurring distress about the event in the last five years were excluded as well.

Six subjects were excluded from analyses due to data quality issues head movement greater than 5mm away from DLPFC target during the rTMS session (for >5 min, cumulatively), leaving 60 subjects for further analysis.

**Study Design**—In our primary experiment, we employed a 2×2 between-groups design, with rTMS condition (active vs. sham) and hemisphere (left vs. right DLPFC) as between-subject factors. The 66 participants were randomly assigned into the four groups. After exclusion of the 6 subjects with excessive head movement, that left 14 and 16 subjects in active left and right conditions (respectively), and 13 and 17 subjects in the sham left and right conditions (respectively). Following a screening visit and a structural MRI scan (for active condition participants; see below), all subjects participated in two separate rTMS sessions. We employed a two-session approach owing to the temporal dynamics of 1 Hz rTMS stimulation. Low frequency rTMS has been shown to suppress cortical excitability of the targeted region following stimulation(Robertson et al., 2003), with cognitive and behavioral effects lasting for approximately half the time of stimulation duration(Sandrini et al., 2011; Thut and Pascual-Leone, 2010). The maximum stimulation duration in any one session was approximately 30 min, constraining each of the two rating sessions to 15 minutes. Participants evaluated blameworthiness in one session and assigned punishment in the other. Task order (punishment vs. blameworthiness) was counterbalanced across participants. Participants received the same type of stimulation, to the same hemisphere, in both sessions. The two sessions were separated by no less than 48 hours and no more than two weeks.

**MRI**—For participants recruited into an active condition of the study, we obtained a structural MRI to aid anatomical localization of the DLPFC ROI for rTMS. A T1-weighted high-resolution 3D anatomical scan was obtained for each participant (FOV 256×256, 1×1×1mm resolution). In addition, fast spin echo axial spin density weighted (TE=19, TR=5000, 3 mm thick) and T2-weighted (TE=106, TR=5000, 3 mm thick) slices were obtained to exclude any participants with structural abnormalities that would impede locations. All MRI scans were performed on a 3 Tesla Phillips Achieva scanner located at the Vanderbilt University Institute for Imaging Science (VUIIS).

**rTMS Sessions**—At the start of each experimental session subjects completed the Mini-Mental Status Exam and the TMS/rTMS Acute Side Effects questionnaire (See Supplementary Materials) to obtain baseline measurements for comparison with post-scan ratings. The experimenter explained the task instructions to participants, who completed 5 practice trials for the session-appropriate judgment type (i.e. blameworthiness vs. punishment). The practice scenarios spanned the full extent of crime severity and responsibility of the scenarios used in the experimental session in order for subjects to calibrate their ratings along the entire 10-point Likert scale.
rTMS stimulation was then applied for 30 minutes to left or right DLPFC using a Magstim 2T Rapid stimulator (30% of maximum output), and either MagStim placebo coil (Sham condition) or a MagStim 70-mm air-cooled figure-8 coil (Active condition). rTMS pulses were triggered remotely using a computer. Sham stimulation produced a click that resembled the sound of rTMS, however no magnetic pulse was delivered. For participants in the active condition of the study, DLPFC was localized for manual targeting using theBrainsight frameless stereotaxic system (Rogue Research, Toronto CA), which was calibrated prior to each session. Participants' structural MRIs were warped to a common coordinate space, and the DLPFC target set on each subject's Talairach-warped brain surface. Talairach coordinates were +/-39, 37, 22 [x,y,z], corresponding to left or right Brodmann area 46. This coordinate was chosen because it was the focus of peak activation for the DLPFC region engaged by punishment decisions in our prior study (Buckholtz et al., 2008). Participants were seated in a comfortable lounger with custom adaptations to integrate (and provide ergonomic support for) the stereotactic apparatus. Once the participant was situated comfortably and securely for frameless stereotaxy, the experimenter manually positioned the coil until it was within 3mm of the target; after positioning, the coil was locked into place. Target localization was continuously monitored throughout the stimulation session to detect head movement (especially, movement that repositioned the participant's stimulation target >3mm from the coil, as indicated via Brainsight). For participants in the sham condition of the study, the rTMS coil was positioned in a spot approximating the stimulation coordinate for DLPFC. While frameless stereotaxy was not employed for these individuals, the experimenters performed a "mock" localization and positioning to increase believability. Immediately following stimulation, participants were asked to perform the rating task (i.e. punishment or blameworthiness assessment) on a computer that was directly adjacent to the rTMS apparatus. After participants completed the task, we again administered the TMS/rTMS Acute Side Effects questionnaire and Mini-mental status exam to monitor for any adverse effects of rTMS.

**Experimental Paradigm**—The scenarios used in the present study were the same as in Buckholtz et al. (2008) (Buckholtz et al., 2008). On each trial of the task, subjects were shown a short written scenario depicting the actions of a protagonist named “John.” These scenarios were divided into two conditions: Responsibility and Diminished-Responsibility. In the Responsibility condition (R), John was described committing a crime, ranging from simple theft to assault and murder. In the Diminished-Responsibility (DR) condition, John was described committing crimes of similar magnitude, but these scenarios also contained mitigating, justifying, or otherwise excusing circumstances that reduced John's level of criminal culpability.

As in Buckholtz et al. (2008), the R and DR scenarios were variants of one another (though no subject saw both variants of the same scenario). We constructed 2 sets of scenarios, each with the same scenario “stem.” The Responsibility scenarios of “set 2” consisted of “set 1” Diminished-Responsibility scenarios from which the mitigating circumstances had been excised, while the Diminished-Responsibility scenarios of set 2 consisted of set 1 Responsibility scenarios with mitigating circumstances added. The assignment of scenario set to judgment type (blameworthiness vs. punishment) was counterbalanced across subjects.
(e.g. subject 1 received set 1 scenarios for blameworthiness ratings and set 2 scenarios for punishment ratings, while subject 2 received set 1 scenarios for punishment ratings and set 2 scenarios for blameworthiness ratings, and so forth). Thus, exactly the same premises were used in constructing the R and DR scenarios. Piloting confirmed that neither blameworthiness nor punishment ratings differed between the different scenario sets.

In each session, subjects were asked to make ratings on 28 such scenarios (14 R, 14 DR). During the “punishment” session, subjects were asked to “Please indicate how much punishment John deserves for his actions described in the scenario, on a Likert scale from 0-9, where 0 = No Punishment and 9 = Extreme Punishment.” During the “blameworthiness” session, subjects were asked to “Please indicate how morally responsible John is for his actions described in the scenario, on a scale from 0-9, where 0 = Not Morally Responsible At All and 9 = Completely Morally Responsible.” Participants were asked to consider each scenario (and thus, each “John”) independently of the others and were encouraged to use the full scale (0–9) for their ratings. In the scanner but prior to the functional scans, subjects were shown five practice scenarios that were designed to span the punishment scale. Scenarios were presented as white text (Times New Roman font) on a black background (14.2 degrees [width]× 9.9 degrees [height] of visual angle). Below each scenario, text reminded participants of the task instructions. Participants made their ratings via keypress using a standard keyboard (i.e. buttons 0-9). The rating tasks were self-paced, with each scenario presented until the participants made a response, or until a maximum of 40s. Participants were instructed to make a response as soon as they had reached a decision. Total maximum possible session duration was 21 minutes; however, average session duration was 14 minutes. The experiment was programmed in Matlab (Mathworks, Natick MA) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997) and was presented using a Pentium IV PC.

**Mediation analysis**—We used conditional process analyses to test the hypothesis that DLPFC rTMS affects punishment ratings by modulating the influence of culpability and harm on punishment decisions (Hayes, 2013). These analyses were performed using the SPSS macro “PROCESS,” made available by Andrew Hayes(Hayes, 2013). We constructed a conditional process model that specified rTMS group (Active or Sham) as an independent variable, $\beta_{\text{Culpability}}$ and $\beta_{\text{Harm}}$ values as mediators and punishment values as a dependent variable. Punishment set and gender were included as covariates. We used a non-parametric resampling procedure (bootstrapping) with 5000 samples to estimate the significance of the indirect effect. Through bootstrapping, we calculated point estimates of the indirect effects for each mediator over all samples, and constructed a 95% confidence interval around each point estimate. Statistical significance is inferred if the upper and lower bounds of the confidence interval do not contain zero. Bootstrapping is generally considered preferable to parametric tests of mediation (e.g. the Sobel test), because it avoids the assumption of normality for the sampling distributions of the total and specific indirect effects, which is typically violated in practice.
fMRI Study

Participants—Ten healthy community volunteers were recruited to participate in this study (aged 18-36; 7 males). All participants provided written informed consent, and all procedures were approved by the Vanderbilt University Institutional Review Board (IRB). Exclusionary criteria were identical to those used in the rTMS portion of this study, described above.

Experimental Stimuli—Stimuli and ratings were the same as those used in the rTMS experiment, described above, with the exception that in addition to rating punishment and blameworthiness, participants also rated how long it took John to plan and execute the actions described in the scenario. Results from this rating will not be described here, as these data were collected as part of a separate experiment. Participants saw 84 of these vignettes, split evenly between the R and DR conditions and between the different judgment types. Importantly, the number of R and DR conditions was identical across the judgment types, and the order of judgment type was counterbalanced across subjects.

Experimental Protocol—Subjects were asked to make one of the 3 ratings described above while undergoing an fMRI scan. Participants completed 12 runs, with each rating completed in a block of 4 sequential runs. At the beginning of each block, before the scanner started, subjects completed 5 practice trials to familiarize them with the rating for that block. Each run consisted of 7 trials split between the R and DR conditions, with a minimum of 1 trial of each type presented in each run. Scenarios were presented during scanning and during practice using a visual display presented on an LCD panel and back-projected onto a screen positioned at the front of the magnet bore. Subjects were positioned supine in the scanner so as to be able to view the projector display using a mirror above their eyes. Manual responses were recorded using two five-button keypads (one for each hand; Rowland Institute of Science, Cambridge, MA). Subjects were instructed to make a manual response as soon as they had arrived at a decision, so as to ensure that neural activity around the time of response would reflect decision-making. After each manual-press, subjects viewed a fixation cue for a 12 s inter-trial interval (ITI). Subjects had up to 45 seconds to respond on each trial. If this limit was reached without a response, the trial automatically moved into the ITI.

fMRI Data Acquisition—All fMRI scans were acquired using a 3T Philips Achieva scanner at the Vanderbilt University Institute of Imaging Science. Stimulus presentation was synchronized to fMRI volume acquisition. Functional (T2* weighted) images were acquired using a gradient-echo echoplanar imaging (EPI) pulse sequence with the following parameters: TR = 2000 ms, TE = 35 ms, flip angle 79°, FOV 192 x 192 mm, 64 x 64 matrix with 34 axial slices (3.0 mm, 0.3 mm gap) oriented at a 15° oblique angle to the AC-PC. As the length of each trial – and therefore each run – was subject-dependent, run length was adjusted for each scan based on that subject's speed on the previous run. Runs ranged from 3 minutes 30 seconds to 6 minutes 40 seconds depending on the speed of subjects' responses.

Preprocessing—Image analysis was conducted using Brain Voyager QX 2.3 (Brain Innovation, Maastricht, The Netherlands) in conjunction with custom Matlab software. All
images were preprocessed using standard 3D motion correction, slice timing correction, linear trend removal, and spatial smoothing with a 6 mm Gaussian kernel (full width at half maximum) as implemented through Brain Voyager software. Subjects’ functional data were co-registered with their T1-weighted anatomical volumes and transformed into standardized Talairach space.

**Statistical Analysis**—A random-effects general linear model (GLM) was constructed by convolving a canonical hemodynamic response function (double gamma, including a positive $\gamma$ function and a smaller, negative $\gamma$ function to reflect the BOLD undershoot) to the following set of regressors: a ‘decision’ epoch starting at the time-point 3 TRs (6 seconds) prior to each response and ending with the TR in which the response was made, and a baseline that included all other epochs. The selection of the decision epoch was motivated by the expectation that decision-related modulation of BOLD signal would correspond with the portion of the timecourse prior to and up to the participants’ responses (Cushman et al., 2012; Shenhav and Greene, 2010). Beta-weights for each fMRI run were transformed into Z-scores signifying the magnitude of deviation of fMRI signal during the decision-epoch as compared to the average signal during all other periods.

**A Priori ROI Definition**—We examined timecourse data from our two *a priori* DLPFC ROIs. Both ROIs were defined by a 5mm cube centered on the peak coordinate also targeted by rTMS, identified in our previous study (Buckholtz et al., 2008). Timecourses of activation were created for each condition (R and DR) for each judgment type (Punishment and Responsibility) for each subject by collapsing across epochs for each cell of this $2 \times 2$ design for each subject. Timecourses were compared to a baseline of the overall average for that condition for that rating.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

We wish to thank Brian Essex for assistance with the rTMS localization protocol, Randolph Blake for kindly making available TMS equipment, and Ashley Schwartzman for assistance with rTMS sessions. We are grateful to Fiery Cushman, Liane Young, and Joshua Green for comments on this manuscript. This study was supported by a grant from the John D. and Catherine T. MacArthur Foundation to Vanderbilt University and the University of California at Santa Barbara. Its contents do not necessarily represent the official views of either the John D. and Catherine T. MacArthur Foundation or the MacArthur Foundation Research Network on Law and Neuroscience (www.lawneuro.org). JWB was supported by the NIMH (T32MH018921 and T32MH064913), the NIDA, the Sloan Foundation, the Brain and Behavior Research Foundation, and the MGH Center for Law, Brain, and Behavior.

**References**


Figure 1. Harm and Culpability Affect Blameworthiness and Punishment Decisions
Mean ratings of Blameworthiness (A) and Punishment (B) as a function of Harm severity (x-axis) and Culpability (colored lines). (C) and (D) depict mean ratings across all combinations of Culpability and Harm, ordered from low Culpability/low Harm to full Culpability/high Harm. Error bars indicate the standard error of the mean (SEM).
**Fig 2. DLPFC Stimulation Site**

DLPFC (Talairach +/-39, 37, 22 [x,y,z]) was localized for each subject by warping individual structural MRI's to the Talairach template. (A) Trajectory and approach angle (green funnel) calculated by Brainsight to guide coil placement for DLPFC target coordinate (red dot). Trajectory and target are visualized on a 3-dimensional curvilinear surface reconstruction of one individual participant's warped T1 MRI. (B) Location of the DLPFC target (red dot) on an individual participant's warped T1 image (L=R, R=L). Green crosshair indicates skull contact point for stimulation coil.
Fig 3. TMS Selectively Affects Punishment
Mean Blameworthiness (A) and Punishment (B) z-scores as a function of TMS stimulation condition (Active vs. Sham) and Culpability (colored error bars). Given that the ratings for R and DR trials occupied different portions of the scale, we z-transformed the mean ratings to emphasize the difference in rTMS effects between each of the Punishment and Blameworthiness conditions. (C) and (D) The specific ‘locus’ of the differential effect of rTMS on Punishment and Blameworthiness ratings is revealed by plotting the mean ratings across all combinations of Culpability and Harm, ordered from low Culpability/low Harm to full Culpability/high Harm. Error bars indicate the standard error of the mean (SEM).
Fig 4. Relationship between, and rTMS effects on, Harm and Culpability Betas for Punishment Decisions

(A) Negative correlation between beta-weights derived from linear regression models with Harm Severity and Culpability Level as predictors. Values shown were obtained by z-transforming the absolute value of beta-weights for each predictor. Separate regression models were created for each participant to create per-subject beta-weights. (B) Impact of DLPFC rTMS on Harm and Culpability beta-weights.
Fig 5. DLPFC rTMS affects Punishment by Disrupting The Use of Harm and Culpability Signals

Mediation analysis depicts coefficients and standard error (italics) for the effect of rTMS on Harm and Culpability β-weights (A; culpability/harm), and the impact of these β-weights on punishment during R trials (B). Coefficients are standardized, with sign indicating the direction of the relationships. For example: path (A) indicates that DLPFC rTMS decreases the influence of harm information on punishment, and path (B) reveals that culpability betas are positively correlated with punishment. Harm-related coefficients in red, culpability-related coefficients in blue. Path C shows the total effect of TMS on punishment; path C’ shows the direct effect of TMS on punishment (dashed line). Point estimates of indirect effects for both Harm and Blame signals both fall within a 95% confidence interval that does not cross zero, unlike the direct effect of rTMS on punishment (see Results).
Fig 6. DLPFC Activity Selective for Punishment
(A) Depicts 5mm sphere around right DLPFC stimulation target coordinate, from which task- evoked BOLD signal was extracted for condition comparisons. (B) shows right DLPFC BOLD percent signal change from baseline as a function of judgment type (x-axis) and culpability (colored lines).