COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12838

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12838

Moral Values Reveal the Causality Implicit in Verb Meaning

Laura Niemi,[a,b] Joshua Hartshorne,[c] Tobias Gerstenberg,[d] Matthew Stanley,[e] Liane Young[c]

[a]*Department of Psychology, Cornell University*
[b]*Munk School of Global Affairs & Public Policy, University of Toronto*
[c]*Psychology Department, Boston College*
[d]*Department of Psychology, Stanford University*
[e]*Duke Institute of Brain Sciences, Duke University*

Received 26 March 2019; received in revised form 17 March 2020; accepted 6 April 2020

Abstract

Prior work has found that moral values that build and bind groups—that is, the binding values of ingroup loyalty, respect for authority, and preservation of purity—are linked to blaming people who have been harmed. The present research investigated whether people's endorsement of binding values predicts their assignment of the causal locus of harmful events to the victims of the events. We used an implicit causality task from psycholinguistics in which participants read a sentence in the form "SUBJECT verbed OBJECT because…" where male and female proper names occupy the SUBJECT and OBJECT position. The participants were asked to predict the pronoun that follows "because"—the referent to the subject or object—which indicates their intuition about the likely cause of the event. We also collected explicit judgments of causal contributions and measured participants' moral values to investigate the relationship between moral values and the causal interpretation of events. Using two verb sets and two independent replications (N = 459, 249, 788), we found that greater endorsement of binding values was associated with a higher likelihood of selecting the object as the cause for harmful events in the implicit causality task, a result consistent with, and supportive of, previous moral psychological work on victim blaming. Endorsement of binding values also predicted explicit causal attributions to victims. Overall, these findings indicate that moral values that support the group rather than the individual reliably predict that people shift the causal locus of harmful events to those affected by the harms.

Keywords: Morality; Social cognition; Causal attribution; Implicit causality; Psycholinguistics; Moral psychology

Correspondence should be sent to Laura Niemi, Cornell University, Uris Hall, 109 Tower Rd, Ithaca, NY 14853. E-mail: niemi@cornell.edu

1. Introduction

Some moral values protect individuals from harm, whereas other values serve to keep groups intact. At the same time, regardless of moral perspective, harm-doers are typically thought to be causally and morally responsible for their wrongdoing, and recipients of harm are not. However, moral judgment does not always follow this template—sometimes victims are deemed responsible and blameworthy. The current research investigates the possibility that moral values that prioritize groups over individuals increase the likelihood that harmful events are attributed to the people affected by them.

Across the political spectrum and around the world, when people are asked how they determine what is morally right and wrong, there is some notable consistency across individuals. People commonly consider the presence of harm and injustice to be highly morally relevant. Findings from research inspired by the moral foundations theory demonstrate that *caring* and *fairness* values, which reflect concerns for harm and injustice, are highly endorsed around the world (Graham, Haidt, & Nosek, 2009; Graham et al., 2011; Haidt, 2001, 2007). These values are often called "individualizing values" because they concern the well-being of each individual, regardless of group membership (Graham et al., 2009, 2011).

By contrast, there is more variability regarding the endorsement of values concerned with building and binding groups. These "binding values" concern *loyalty* to the ingroup, obedience to and respect for *authority*, and preservation of *purity* (Graham et al., 2011). For example, politically conservative, lower socioeconomic status, more Machiavellian, and non-WEIRD (Western, Educated, Industrialized, Rich, Democratic; Henrich, Heine, & Norenzayan, 2010) participants tend to endorse binding values more than Western, educated, liberal participants (e.g., Graham et al., 2009, 2011; Haidt, Koller & Dias, 1993; Niemi & Young, 2013, 2016). When asked about what criteria affect their judgments of right and wrong, most people agree about the importance of individualizing values, prohibiting intentional harm and unfairness. Yet there is substantial variability across individuals in their endorsement of binding values (Graham et al., 2009, 2011).

The expanded system of moral values introduced by moral foundations theory allows for a better understanding of the world's diverse people and cultures (Haidt, 2007). However, it is unclear how to reconcile moral foundations theory with theories of moral cognition and judgment centered around prohibiting intentional causation of harm. On a different theoretical account of moral cognition, the dyadic morality framework (Gray, Young, & Waytz, 2012), researchers present the paradigmatic form of immoral events as the "moral dyad," where AGENT-HARMS-PATIENT. According to dyadic morality, moral transgressions are events in which an agent harmed a patient. The account offers a formula for moral judgments that coheres with individualizing values: Agents, the doers of harm, are causal, responsible, and blameworthy, whereas patients, the recipients of harm, are not.

Interpretation of causation in the dyadic morality account (Gray et al., 2012) overlaps with the interpretation of causation (and terminology) surrounding harm in linguistic theory. A widely used classification system of English verbs, which groups together verbs

by their shared syntactic and semantic features (Kipper, Korhonen, Ryan, & Palmer, 2008; Levin, 1993), attaches a "CAUSE" property to the "agent," the initiator of the event, who was intentional or conscious of what they were doing, and who exists outside of the context of the event. The affected "patient" is merely assigned properties that indicate the relevant change in bodily state: for example, in the case of killing, the patient was alive at the start of the event, and not alive as a result of the event. Thus, on the linguistic account, a killing event involves a causal, intentional agent, and an affected, non-causal patient.

Endorsement of binding values, however, implies a very different perspective on cause and effect. First, a violation of the binding values of loyalty, obedience to authority, and preservation of purity may occur in the absence of obvious harm by an agent to a patient. For example, violations can occur without clear boundaries between the roles of agent and patient. In the case of purity violations, the victim can also be seen as the agent (e.g., "tainted" rape victims who are blamed based on purity norms; Chakroff & Young, 2015; Niemi & Young, 2016). Moreover, by definition, the same event ("*A killed B*") that is immoral from the perspective of someone who endorses individualizing values might be morally obligatory from the perspective of someone else who endorses binding values ("*A was ordered by a superior to kill the traitor B*").

Previously, Niemi and Young (2016) found that people who scored higher in binding values blamed patients more and held them more responsible. The relationship between binding values and blame of patients was mediated by the extent to which participants held patients responsible. These results suggest that endorsing binding values might be associated with a different understanding of dyadic harm. People higher in binding values consider moral violations of disloyalty, disobedience, and impurity to be morally relevant —these violations are less likely to have a single, clear causal agent and, in some cases, may be understood as caused by the patient. Thus, people higher in binding values may be more likely to understand harm events encountered in this research as caused not by the agent but by the affected patient. The present research investigates whether people who endorse binding values are more likely to construe harm events as having been caused by the victim.

2. Novel pathways to gain insight on causality and morality

We investigate the hypothesis that people higher in binding values are more likely to shift the causal locus from agents to patients with the Implicit Causality task, a task from psycholinguistics that measures intuitions about likely causes for events (Brown & Fish, 1983; Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Rudolph & Forsterling, 1997).

In this task, participants read prompts such as

(a) Bob murdered Amy because...

and select the word they think follows: the pronoun referring to the agent ("he") or the patient ("she"). Participants' selections reveal an expectation that the murder is due to something that either the agent (subject; Bob) did or the patient (object; Amy) did. People's implicit causality selections provide a window into how they understand the cause of an event: Is the causal locus perceived to reside with the agent or the patient (e.g., Hesslow, 1988; Hilton, 1990; Lombrozo, 2006)? Selecting Bob ("he") for (a) "*Bob murdered Amy because…*" is more consistent with the moral dyad framework, in which the causal locus of harm is placed on agents and not on affected patients (Gray et al., 2012). In contrast, selecting Amy ("she") would be inconsistent with the moral dyad framework, but aligns with the hypothesis that people higher in binding values shift the causal locus from harm-doers to harm-recipients.

The implicit causality task derives its name from the causality implicit in verb meaning. Prior psycholinguistics research (Hartshorne, 2013; Hartshorne & Snedeker, 2012; Kipper et al., 2008) has demonstrated that implicit causality responses tend to vary by verb class. Some verbs reliably prompt selections of pronouns that refer to the subject (e.g., "subject-biased" verbs like *frighten, amuse*); some prompt selections of pronouns that refer to the object (e.g., "object-biased" verbs like *praise, thank*), suggesting that people have systematic expectations about how some categories of events came about (Bott & Solstad, 2014; Brown & Fish, 1983; Ferstl, Garnham, & Manouilidou, 2011; Garvey & Caramazza, 1974; Hartshorne, 2013; Hartshorne & Snedeker, 2012; Kipper-Schuler, 2006; Pickering & Majid, 2007; Rudolph, 2008; Rudolph & Forsterling, 1997). Some other work has shown that implicit causality responses are affected by perceived social hierarchy as well as implied gender roles (Bott & Solstad, 2014; Ferstl et al., 2011; Garvey & Caramazza, 1974; Hartshorne, 2013; Pickering & Majid, 2007).

More recently, research has examined gender differences in pronoun comprehension and interpretation (Arnold, 2015), as well as differences in causal inferences and pronoun expectations for political events (Niemi, Roussos, & Young, 2019; von der Malsburg, Poppels, & Levy, 2018). The evidence so far indicates that implicit causality responses are systematically related to certain individual differences and have the potential to be affected by people's moral values as they consider morally relevant events. However, this topic is still very much unexplored. It remains unclear whether there is indeed a relationship between selections in the implicit causality task and endorsement of different moral values.

Researchers studying the psychology of language and morality have traced other lexical features through which people convey and modulate moral judgments, largely focusing on character judgments. For example, describing a person's contribution to events with abstract adjectives *versus* action verbs has more profound repercussions on moral judgments. For example, a person who is "*aggressive*" may be perceived to have a more persistent character issue, compared to a person who "*aggressed*," which suggests an isolated event (Fiedler & Krüger, 2014). Likewise, the linguistic ingroup bias (e.g., Maass, Ceccarelli, & Rudin, 1996) involving withholding negative, but not positive, adjective labels when describing members of the ingroup has clear moral relevance. The current work extends and broadens this prior research by examining how individual differences

in people's moral values relate to their general understanding of the causal locus for morally salient events.

By leveraging the implicit causality task in this research, we were able to investigate whether people high in binding values are more likely to complete sentences like "*Bob murdered Amy because*" with reference to the object (the victim). While this prediction followed findings that binding values predict blame and stigmatization of harmed people (Niemi & Young, 2016), less consistent prior evidence had linked individualizing values to blame of the agent. We also tested responses to morally irrelevant, neutral verbs to rule out the possibility that people high in binding values were more likely to consider patients, more generally, to be causal contributors.

We measured participants' explicit causal judgments, including judgments about whether the agent's action was necessary and sufficient for the outcome, and whether the patient allowed, controlled, and deserved what happened. Because violations of binding values are not necessarily relevant to the moral dyad framework, we expected participants high in binding values to be less likely to view agents as necessary or sufficient and more likely to view patients as having allowed, controlled, or deserved the events. We also expected that implicit causality object-bias—increased likelihood of referring to the sentence object following "*Subject verbed object because*"—would be directly related to judgments that agents were less necessary and sufficient and that patients allowed, controlled, and deserved the events.

Finally, we measured participants' sensitivity to suffering (how "injured" participants considered patients) and stigmatization of patients (how "contaminated" participants considered patients) to understand how these explicit morally relevant attitudes about harmed people (sensitivity vs. stigmatization) relate to implicit causality selections. In prior work (Niemi & Young, 2016), participants who were more sensitive to patient suffering—rating patients as more "injured"—also exhibited higher individualizing values. Increased stigmatization of patients—rating patients as more "contaminated"—was associated with higher binding values.

Moral values were measured with the Moral Foundations Questionnaire (see Section 4), which has been used extensively in prior work to measure people's diverse moral values (e.g., Graham et al., 2011; Niemi & Young, 2016).

3. Overview of studies

The present research tests four hypotheses. The first three examine how moral values are related to implicit causality responses, explicit causal judgments, and the propensity to stigmatize (or be sensitive to) patients. The last hypothesis examines their interrelationships.

1: In the implicit causality task, people higher in binding values will be more likely to select the object over the subject ("object-bias") for harm and force events, but not neutral events.

2: For harm and force events, binding values will be related to reduced judgments of the agent's necessity and sufficiency, and increased judgments of the patient's capacity to allow, control, and deserve events.

3: Binding values and stigmatization of patients will be positively correlated and individualizing values and sensitivity to patient suffering will be positively correlated.

4: Implicit causality object-bias for harm and force events will be related to reduced sensitivity to suffering; judgments of agents as less necessary and sufficient; greater stigmatization of patients; and increased judgments that patients allowed, controlled, and deserved harm and force events.

We test all four hypotheses in the following study. We also replicate the findings involving the implicit causality bias in Replication Dataset 1, and replicate the findings involving the implicit causality bias using an expanded set of verbs and a larger sample size in Replication Dataset 2.

The results of these studies will illuminate whether and how causal attributions for harmful events differ when groups are morally prioritized over the individual. If binding values predict increased causal attribution to victims, this will extend the prior research showing a relationship between binding values and explicit victim-blame. Such results would suggest that future models of moral cognition and judgment should take into account the possibility that moral values that support groups systematically alter inferences about causation, increasing the likelihood that harmful events are attributed to the people affected by them.

4. Materials and methods

Participants ($N = 459$) were recruited online via Amazon's Mechanical Turk ($M_{age} = 37.25$ years, $SD_{age} = 31.39$; 207 selected female, 247 selected male, 5 selected other or were missing). We excluded 189 additional individuals who failed attention checks.[1] We aimed to have approximately 200 participants in each condition (*Male-verbed-female versus Female-verbed-male,* described below) in line with past work showing that associations among moral values, blame, and responsibility were found in samples of approximately this size (Niemi & Young, 2016). We also conducted two additional studies to replicate to replicate effects obtained in the primary experiment with Replication Dataset 1 ($N = 249$; $M_{age} = 35.87$ years, $SD_{age} = 13.49$; 114 selected female, 133 selected male, 2 selected other or were missing) and Replication Dataset 2 ($N = 788$; $M_{age} = 36.32$ years, $SD_{age} = 12.88$; 279 selected female, 504 selected male, 5 selected other or were missing).[2] Methodological differences between Study 1 and the Replication Datasets are described in the Supplementary Materials. The institutional ethics review

board approved all studies, and informed consent was obtained from all participants via an online form. Materials used are described in the following sections.

4.1. Moral values

Moral values in the five foundations (caring, fairness, loyalty, obedience to authority, and purity) were assessed using the 30-item Moral Foundations Questionnaire (MFQ-30; Graham et al., 2011, see Appendix). The Moral Foundations Questionnaire consists of two sections in which participants' values in five foundations are determined from their responses to a series of questions. In one section, the prompt is, "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?" and participants respond using a scale from 1 "Not at all relevant" to 6 "Extremely relevant." Items gauging participants' valuation of *authority* in this section include, for example, "Whether or not someone showed a lack of respect for authority"; *loyalty* includes, "Whether or not someone did something to betray his or her group"; *purity* includes, "Whether or not someone violated standards of purity and decency." A different section of the Moral Foundations Questionnaire consists of full sentences to which participants indicate their agreement or disagreement on a Likert scale ranging from 1 to 6. Items gauging participants' endorsement of *caring* values include "Compassion for those who are suffering is the most crucial virtue," and for *fairness* values: "Justice is the most important requirement for a society." "Individualizing values" were computed as the average scores for caring and fairness values; "binding values" were computed as the average scores for loyalty, respect for authority, and purity values. Binding values scores represent the extent to which the participant agrees with or endorses a number of statements which belong to a category of norms concerned with the group, rather than the individual—namely, loyalty, respect for authority, and purity.

Participants also provided demographic information including political orientation, gender, and religiosity, and they completed the Ambivalent Sexism Inventory (Glick & Fiske, 1996; the present analyses do not involve the Ambivalent Sexism Inventory).

4.2. Implicit causality

The implicit causality task involved 24 minimal event descriptions in the form: "[*Subject*] *verb-ed* [*Object*] *because*...," for example, *"Bob coerced Amy because...,"* with half of the participants receiving male sentence subjects and female sentence objects, and vice versa for the other half to equalize gender of the person in the subject and object positions. Participants were asked to "Please select which word you think would follow." They were offered the choices "*he*" or "*she*" (counterbalanced order across items). Verbs described highly morally relevant events (the "harm/force" verbs, henceforth) or neutral events ("neutral filler" verbs; see Table 1 for verbs).[3] Note that we purposefully selected verbs describing events likely to be of importance for informing theories of morality (e.g., *kill* and *rape*). We conducted analyses in which we examined the effect of harm/force verbs and neutral filler verbs using linear mixed-effects models.

Table 1
Verbs and implicit causality biases in Study 1, and Replications 1 and 2

	Study 1	Rep 1	Rep 2
Harm/Force Mean:	0.39	0.41	0.38
Clobbered	0.55	0.57	0.66
Coerced	0.30	0.36	0.27
Enslaved	0.36	0.40	0.24
Forced	0.39	0.37	0.34
Influenced	0.31	0.31	0.26
Killed	0.53	0.57	0.50
Manipulated	0.29	0.32	0.25
Raped	0.30	0.31	0.19
Robbed	0.26	0.28	0.31
Stabbed	0.50	0.54	0.41
Strangled	0.54	0.55	0.44
Tempted	0.31	0.33	0.26
Assaulted			0.41
Convinced			0.35
Enticed			0.30
Groped			0.32
Impaled			0.65
Molested			0.22
Persuaded			0.34
Pressured			0.37
Seduced			0.27
Silenced			0.68
Spanked			0.66
Walloped			0.66
Neutral Fillers Mean:	0.61	0.61	0.58
Approached	0.39	0.41	0.35
Congratulated	0.89	0.88	0.91
Delighted	0.33	0.34	0.34
Impressed	0.21	0.25	0.21
Observed	0.60	0.64	0.58
Praised	0.85	0.85	0.89
Quoted	0.65	0.61	0.68
Skipped	0.57	0.59	0.54
Thanked	0.85	0.86	0.86
Transported	0.77	0.71	0.81
Boggled			0.36
Caressed			0.40
Celebrated			0.84
Comforted			0.82
Appraised			0.65
Complimented			0.82
Honored			0.86
Massaged			0.61
Diverted			0.49

(continued)

Table 1. (continued)

Study 1	Rep 1	Rep 2
Fondled		0.26
Greeted		0.51
Puzzled		0.23
Tickled		0.52
Raced		0.33

Note. Higher value indicates greater object-bias in the implicit causality task. Selection of the referent to the object is coded as 1 and subject as 0.

4.3. Explicit causality

To assess explicit (as opposed to implicit) causal judgments, we collected participants' judgments about agents' and patients' causal contributions. Because these items measure perceived causal contribution as a number of scalar dimensions—rather than one forced-choice (X or Y) selection—we were able to investigate whether people indeed ascribe these dimensions of causation predominantly to agents, favoring the moral dyad theory, or whether people shift causation to patients, or whether people assign causation to both agents and patients alike. After completing all the implicit causality task items, participants viewed the same events they had seen in the implicit causality block without the "because" connective (e.g., *"Bob coerced Amy."*). They were asked to "weigh the following possibilities" in the following order (responses were collected using sliding scales, 0 = "Definitely No," 50 = "Unsure," 100 = "Definitely Yes"):

1. Agent Unnecessary: for example, "Would *[Amy]* have been *[coerced]* by someone else?"
2. Agent Sufficient: for example, "Would *[Bob] [coerce]* someone else?"
3. Patient Control: for example, "Did *[Amy]* have control over the occurrence of the event?"
4. Patient Allowing: for example, "Did *[Amy]* let the event happen?"
5. Patient Desert: for example, "Could *[Amy]* have deserved the event?"

4.4. Sensitivity versus stigmatization

Finally, we measured sensitivity to suffering versus stigmatization of patients as in prior work (Niemi & Young, 2016). We asked participants to rate in counterbalanced order how "*contaminated/ tainted*" and "*injured/ wounded*" they considered hypothetical crime victims (crimes: *molestation*, *rape, strangling, stabbing*) on a sliding scale from 0 (*Not at all*) to 7 (*Very much*). As in Niemi and Young (2016), only these four crimes were used to obtain measures of how "*contaminated/ tainted*" and "*injured/ wounded*" participants rated hypothetical crime victims. Average ratings across events as contaminated/tainted and injured/wounded served as indices of stigmatization and sensitivity to suffering, respectively.

4.5. Statistical analyses

Data were analyzed in several ways to address different questions. First, using R (R Development Core Team, 2009) with the lme4 software package (Bates, Maechler, Bolker, & Walker, 2015), we computed a series of generalized linear mixed-effects models. For models with binary outcome variables, significance and 95% CIs around beta-estimates were computed using Wald tests. For models with non-binary outcome variables, significance for fixed effects was assessed using Satterthwaite approximations to degrees of freedom, and 95% CIs around beta-estimates were computed using parametric bootstrapping. In all models, participant and verb were included as crossed random effects (random intercepts only).[4] Finally, to address questions about the relationship between moral values, stigmatization, and sensitivity to suffering, we computed a series of Spearman's rank-order correlations.

5. Results

We address our four hypotheses in the order that they were presented in the Introduction.

5.1. Moral values and implicit causality object-bias

We expected those higher in binding values to be more likely to select the object over the subject as the referent ("object-bias") for harm and force events, but not for neutral events. We tested the relationship between moral values and the implicit causality object-bias with a series of generalized linear mixed-effects models (link = "logit"). First, a generalized linear mixed-effects regression model was computed in which verb type (harm/force (coded as 0) *versus* neutral filler (coded as 1)) and binding values were included as fixed predictors of the propensity to select the object (coded as 1) *relative to* the subject (coded as 0) as the referent. There was a significant interaction between verb type and binding values in Study 1 and in the Replication Datasets (see Table 2). To further interrogate these significant interaction effects, follow-up generalized linear mixed-effects models were computed for harm/force verbs and for neutral filler verbs, taken separately.

For harm/force verbs only, a generalized linear mixed-effects regression model was computed for which binding values were included as the fixed predictor of the propensity to select the object (coded as 1) *relative to* the subject (coded as 0) as the referent. This analysis yielded a significant relationship between binding values and the likelihood of selecting the object as the referent ($b = 0.40$, $SE = 0.06$, $Z = 6.51$, $p < .0001$, odds ratio = 1.50, 95% CI = [0.28, 0.52]). We obtained the same pattern of results in Replication Dataset 1 ($b = 0.50$, $SE = 0.10$, $Z = 4.79$, $p < .0001$, odds ratio = 1.64, 95% CI = [0.29, 0.70]), and Replication Dataset 2 ($b = 0.22$, $SE = 0.06$, $Z = 4.00$, $p < .0001$, odds ratio = 1.25, 95% CI = [0.11, 0.33]). In all three datasets, participants higher in binding values were more likely to select the object over the subject as the referent (object-bias) for harm and force events (see Fig. 1).

Table 2
The results of generalized linear mixed-effects regression models—each with verb type and binding values as predictors of selecting object relative to the subject as the referent

	b	*SE*	*Z*	*p*	95% CI	Odds ratio
Study 1						
Verb Type	2.17	.43	5.06	<.0001	[1.33, 3.02]	8.79
Binding Values	0.37	0.05	7.02	<.0001	[.27, 0.48]	1.45
Verb Type × Binding Values	−0.27	0.05	−5.60	<.0001	[−.36, −0.18]	.76
Replication Dataset 1						
Verb Type	2.43	.47	5.17	<.0001	[1.51, 3.36]	11.40
Binding Values	0.43	0.08	5.27	<.0001	[.27, .59]	1.54
Verb Type × Binding Values	−0.35	0.07	−4.76	<.0001	[−0.49, −0.20]	.71
Replication Dataset 2						
Verb Type	1.87	0.47	4.00	<.0001	[0.95, 2.79]	6.51
Binding Values	0.19	0.04	4.37	<.0001	[0.11, 0.28]	1.21
Verb Type × Binding Values	−0.11	0.03	−2.53	.011	[−0.19, −0.02]	.90

Note. Study 1 (N = 459), Replication Dataset 1 (N = 249), Replication Dataset 2 (N = 788). All 95% CIs are for the beta-estimates. For the "Verb Type" variable, harm/force verbs were coded as 0, and neutral filler verbs were coded as 1.

Importantly, for the neutral filler verbs only, there was no significant relationship between binding values and the selection of the object *relative to* the subject as the referent in the primary study (b = 0.11, SE = 0.06, Z = 1.82, p = .068, odds ratio = 1.11, 95% CI = [−0.01, 0.22]), Replication Dataset 1 (b = 0.08, SE = 0.08, Z = 1.03, p = .31, odds ratio = 1.08, 95% CI = [−0.07, 0.24]), or Replication Dataset 2 (b = 0.08, SE = 0.04, Z = 1.89, p = .059, odds ratio = 1.09, 95% CI = [0.00, 0.17]).

Next, we tested a number of additional considerations related to the implicit causality object-bias. All of these findings are presented in full in Supplementary Materials. First, given that prior work has identified relationships between binding values and political orientation, gender, and religiosity (Graham et al., 2011), we wanted to ensure that binding values predicted the implicit causality object-bias above and beyond these other variables. Our analyses indicate that binding values remain consistent and significant predictors of the implicit causality object-bias for harm/force verbs after statistically controlling for political orientation, gender, and religiosity in all three datasets. Second, we found no relationship between individualizing values and the propensity to select the object *relative to* the subject as the referent for harm/force verbs or neutral filler verbs. Third, gender condition (male-verbed-female *vs.* female-verbed-male) was related to the implicit causality object-bias for harm/force verbs. Specifically, in all three datasets, participants were more likely to select men for harm/force events. However, binding values continued to significantly predict the implicit causality object-bias after statistically accounting for this gender effect.

Overall, the implicit causality results support our first hypothesis. Participants higher in binding values were indeed more likely to select the object over the subject as the

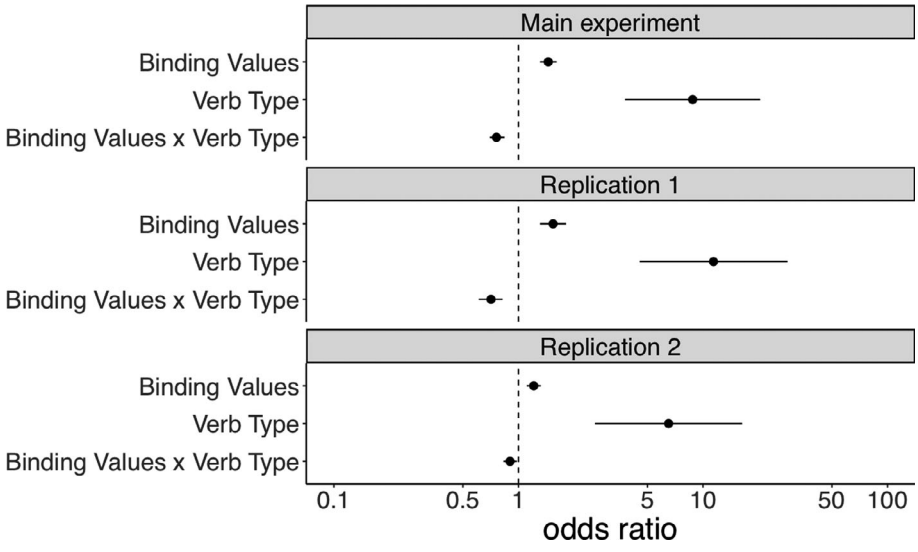

Fig. 1. Plots showing the results of generalized linear mixed-effects regression models—each with verb type and binding values as predictors of selecting object relative to the subject as the referent (error bars represent 95% confidence intervals).

referent for harm and force events, but not for neutral events. These effects remained statistically significant after controlling for a variety of other variables.

5.2. Explicit causal judgments: Agents' and patients' contributions

We next tested our hypothesis that binding values would be negatively related to participants' judgments about agents' necessity and sufficiency, and positively related to their judgments of patients' capacities to allow, control, and deserve events of harm and force; for the purpose of comparison, we also investigated whether the opposite patterns would be observed for individualizing values. We first computed a series of linear mixed-effects models in which binding values and verb type (harm/force [coded as 0] *vs.* neutral filler [coded as 1]) were included as fixed predictors of judgments for necessity, sufficiency, allowing, controlling, and deserving (in separate models). In all five models, there was a significant interaction effect between binding values and verb type (see Table 3 for full results of all models; Fig. 2). To further interrogate these significant interaction effects, follow-up linear mixed-effects models were computed for harm/force verbs and for neutral filler verbs, taken separately.

For harm/force verbs only, we computed five linear mixed-effects models with binding values as the fixed predictor of necessity, sufficiency, allowing, controlling, and deserving judgments (in separate models). These models revealed that binding values were negatively related to participants' judgments about the agent's necessity ($b = -3.36$, $SE = 0.91$, $t = -3.70$, $p = .0002$, 95% CI = [−5.20, −1.73]) and sufficiency ($b = -1.77$, $SE = 0.88$, $t = -2.01$, $p = .046$, 95% CI = [−3.52, 0.13]), and positively related to their

Table 3
The results of five different linear mixed-effects regression models are depicted. In all models, verb type and binding values were fixed predictors; necessity, sufficiency, allowing, controlling, and deserving were the outcome variables in the different models

	b	*SE*	*t*	*p*	95% CI
Outcome: Necessity					
Verb Type	−33.83	4.70	−7.20	<.0001	[−42.77, −25.96]
Binding Values	−3.36	0.78	−4.33	<.0001	[−4.75, −1.86]
Verb Type × Binding Values	2.96	0.40	7.37	<.0001	[2.21, 3.74]
Outcome: Sufficiency					
Verb Type	−3.34	1.97	−1.70	.095	[−7.13, 0.62]
Binding Values	−1.77	0.83	−2.14	.033	[−3.55, −0.16]
Verb Type × Binding Values	1.18	0.35	3.40	.0007	[0.44, 1.87]
Outcome: Allow					
Verb Type	30.95	6.51	4.76	<.0001	[18.59, 44.02]
Binding Values	5.71	0.77	7.42	<.0001	[4.16, 7.37]
Verb Type × Binding Values	−1.71	0.49	−3.52	<.0001	[−2.61, −0.71]
Outcome: Control					
Verb Type	18.30	5.90	3.10	.005	[7.89, 30.81]
Binding Values	4.08	0.74	5.54	<.0001	[2.70, 5.54]
Verb Type × Binding Values	−1.22	0.48	−2.53	.012	[−2.22, −0.39]
Outcome: Deserve					
Verb Type	47.38	5.58	8.49	<.0001	[35.69, 57.98]
Binding Values	3.83	0.66	5.80	<.0001	[2.39, 5.11]
Verb Type × Binding Values	−3.23	0.45	−7.17	<.0001	[−4.04, −2.34]

Note. All 95% CIs are for the beta-estimates. For the "Verb Type" variable, harm/force verbs were coded as 0, and neutral filler verbs were coded as 1.

judgments about the patient's capacity to allow ($b = 5.71$, $SE = 0.88$, $t = 6.52$, $p < .0001$, 95% CI = [3.83, 7.48]), control ($b = 4.08$, $SE = 0.80$, $t = 5.09$, $p < .0001$, 95% CI = [2.55, 5.63]), and deserve ($b = 3.83$, $SE = 0.87$, $t = 4.42$, $p < .0001$, 95% CI = [2.15, 5.52]) the events.

Importantly, for the neutral filler verbs only, there was no significant relationship between binding values and judgments of necessity ($b = -0.40$, $SE = 0.79$, $t = -0.51$, $p = .61$, 95% CI = [−1.95, 1.24]), sufficiency ($b = -0.58$, $SE = 0.85$, $t = -0.69$, $p = 0.49$, 95% CI = [−2.37, 1.05]), or desert ($b = 0.61$, $SE = 0.71$, $t = 0.86$, $p = .39$, 95% CI = [−0.83, 1.98]). However, there were significant and positive relationships between binding values and judgments of allowing ($b = 4.00$, $SE = 0.85$, $t = 4.73$, $p < .0001$, 95% CI = [2.31, 5.63]) and controlling ($b = 2.86$, $SE = 0.83$, $t = 3.45$, $p = .0006$, 95% CI = [1.25, 4.47]). Nevertheless, for allowing and controlling judgments, the magnitude of the effect was larger for harm/force verbs than for neutral filler verbs (see 95% CIs above).

For the purposes of comparison, we next computed a series of linear mixed-effects models in which individualizing values and verb type (harm/force (coded as 0) versus neutral filler (coded as 1)) were included as fixed predictors of judgments for necessity, sufficiency, allowing, controlling, and deserving (in separate models). In models with

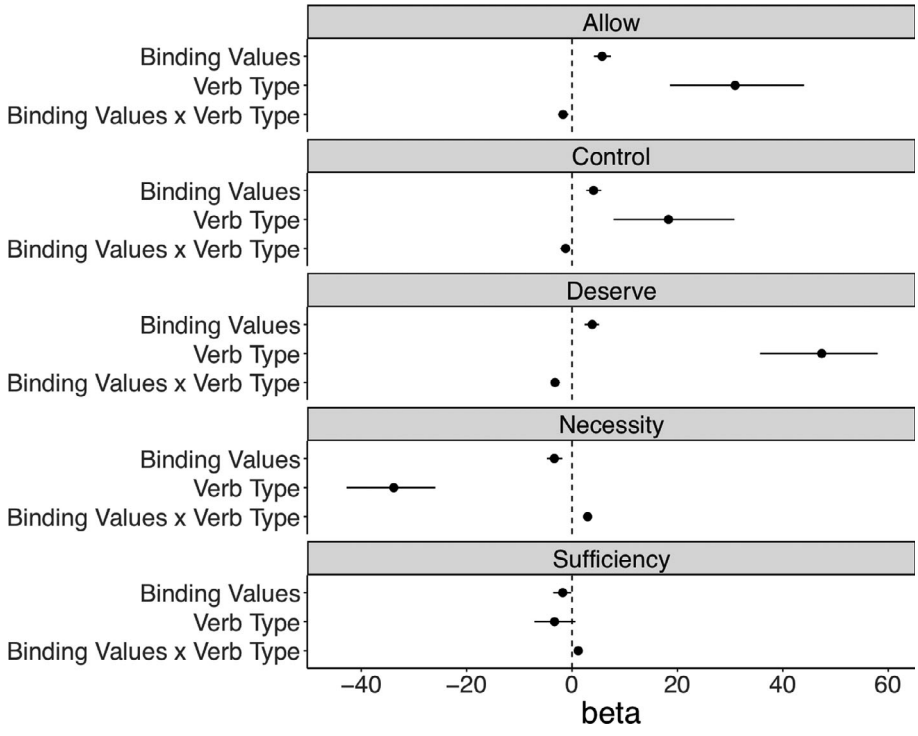

Fig. 2. Plots showing the results of generalized linear mixed-effects regression models—each with verb type and binding values as fixed predictors; allowing, controlling, and deserving, necessity, and sufficiency were the outcome variables in the different models (error bars represent 95% confidence intervals).

necessity, allowing, controlling, and deserving as outcome variables, there were significant interaction effects between individualizing values and verb type (see Table 4 for full results of all models; Fig. 3). For sufficiency judgments, there was only a significant main effect of individualizing values. To further interrogate the four significant interaction effects, linear mixed-effects models were computed for harm/force verbs and for neutral filler verbs, taken separately.

For harm/force verbs only, four linear mixed-effects models with individualizing values as the fixed predictor of necessity, allowing, controlling, and deserving judgments (in separate models). These models revealed that individualizing values were not significantly related to participants' judgments about the agent's necessity ($b = 0.20$, $SE = 1.35$, $t = 0.15$, $p = .88$, 95% CI = [−2.35, 2.78]), but individualizing values were negatively related to judgments of the patient's capacity to allow ($b = -2.65$, $SE = 1.34$, $t = -1.98$, $p = .048$, 95% CI = [−5.24, −0.16]), control ($b = -4.14$, $SE = 1.19$, $t = -3.48$, $p = .0006$, 95% CI = [−6.12, −1.57]), and deserve ($b = -4.16$, $SE = 1.28$, $t = -3.24$, $p = .001$, 95% CI = [−6.80, −1.42]) the events.

For the neutral filler verbs only, there was only a significant (and positive) relationship between individualizing values and judgments of deserving ($b = 2.50$, $SE = 1.03$,

Table 4
The results of five different linear mixed-effects regression models are depicted. In all models, verb type and individualizing values were fixed predictors; necessity, sufficiency, allowing, controlling, and deserving were the different outcome variables in the different models

	b	*SE*	*t*	*p*	95% CI
Outcome: Necessity					
Verb Type	−12.54	5.26	−2.38	.022	[−23.06, −2.04]
Individualizing Values	0.20	1.14	0.18	.859	[−1.93, 2.39]
Verb Type × Individualizing Values	−2.15	0.59	−3.66	.0003	[−3.30, −0.96]
Outcome: Sufficiency					
Verb Type	4.35	2.84	1.53	.127	[−1.20, 9.91]
Individualizing Values	4.64	1.20	3.87	.0001	[2.56, 7.02]
Verb Type × Individualizing Values	−0.69	.51	−1.35	.178	[−1.74, .29]
Outcome: Allow					
Verb Type	11.90	7.10	1.68	.103	[−3.00, 26.27]
Individualizing Values	−2.65	1.18	−2.25	.025	[−4.71, −0.40]
Verb Type × Individualizing Values	2.66	0.71	3.75	.0002	[1.42, 3.89]
Outcome: Control					
Verb Type	4.68	6.55	0.71	.480	[−8.51, 18.00]
Individualizing Values	−4.14	1.09	−3.78	.0002	[−6.21, −2.09]
Verb Type × Individualizing Values	1.91	0.71	2.70	.007	[.56, 3.19]
Outcome: Deserve					
Verb Type	3.68	6.17	0.60	.555	[−9.50, 15.34]
Individualizing Values	−4.16	0.98	−4.24	<.0001	[−6.11, −2.27]
Verb Type × Individualizing Values	6.65	0.66	10.12	< .0001	[5.39, 7.89]

Note. All 95% CIs are for the beta-estimates. For the "Verb Type" variable, harm/force verbs were coded as 0, and neutral filler verbs were coded as 1.

$t = 2.43$, $p = .016$, 95% CI = [0.58, 4.70]). There were no significant relationships between individualizing values and judgments of necessity ($b = -1.95$, $SE = 1.16$, $t = -1.68$, $p = .093$, 95% CI = [−4.42, 0.36]), allowing ($b = 0.02$, $SE = 1.27$, $t = 0.01$, $p = .99$, 95% CI = [−2.51, 2.25]), or controlling ($b = -2.24$, $SE = 1.22$, $t = -1.83$, $p = .069$, 95% CI = [−4.83, 0.22]) for the neutral filler verbs.

Overall, these results support our second hypothesis. For harm/force verbs, binding values were negatively related to participants' explicit causal judgments about the agent's necessity and sufficiency, and positively related to judgments about the patient's capacity to allow, control, and deserve the events. In contrast, for harm/force verbs, an opposing pattern was observed with individualizing values: The endorsement of individualizing values was negatively related to judgments about the patient's capacity to allow, control, and deserve the events.

5.3. Sensitivity versus stigmatization

Regarding the third hypothesis, we expected positive correlations between binding values and stigmatization, and between individualizing values and sensitivity to suffering. First, we computed a series of zero-order correlations to test whether previously observed

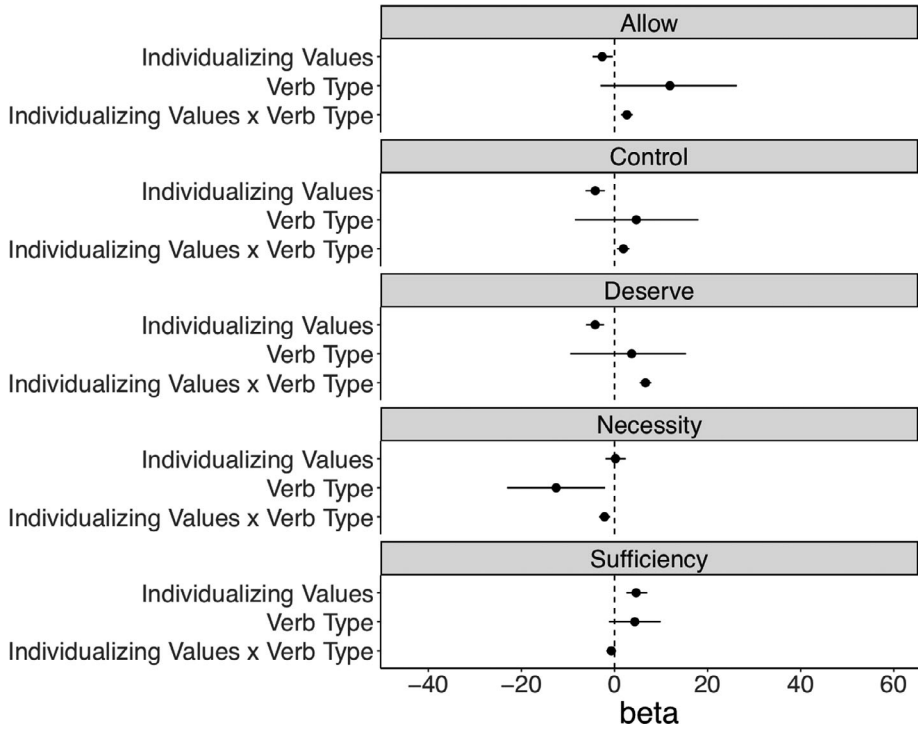

Fig. 3. Plots showing the results of generalized linear mixed-effects regression models—each with verb type and individualizing values as fixed predictors; allowing, controlling, and deserving, necessity, and sufficiency were the outcome variables in the different models (error bars represent 95% confidence intervals).

relationships replicate between binding values and stigmatization, and individualizing values and sensitivity to suffering, for a subset of harmful events (*rape, strangling, stabbing*).[5] We replicated prior findings (Niemi & Young, 2016) of a positive relationship between binding values and ratings of contamination, and a positive relationship between individualizing values and ratings of injury (see Table 5).

Next, regarding the fourth hypothesis, we expected IC object-bias for harm/force events to be related to reduce sensitivity to suffering and greater stigmatization. We also expected implicit causality object-bias for harm/force events to be related to explicit causal judgments (agents' and patients' contributions). We calculated object-bias for harm/force verbs and object-bias for neutral filler verbs by taking the probability of selecting the object as referent across the harm/force events and neutral filler events, respectively. Thus, "harm/force object-bias" represented each participant's tendency toward selecting the object over the subject. Correspondingly, the "neutral filler object-bias" represented a tendency to select the object over the subject across events that do not involve harm/force. Additionally, we created an "*Agent Contribution*" aggregate variable by averaging the agent unnecessary ratings (reverse-coded) and agent sufficiency ratings, and a

Table 5
Spearman's rank-order correlations among moral values, judgments of contamination and injury, implicit causality object-bias for harm and filler events, and explicit causal judgments for agents and patients of harm/force and neutral filler events

							Implicit causality object-bias		Explicit causal ratings		
			Binding	Individ.	Contam.	Injured	Harm	Filler	Harm: Agent	Harm: Patient	Fillers: Agent
Individ.		*Study 1*	0.047								
		Rep 1	0.148*								
		Rep 2	0.139**								
Contam.		*Study 1*	0.473***	−0.101*							
		Rep 1	0.402***	−0.086							
		Rep 2	0.368***	0.053							
Injured		*Study 1*	−0.183***	0.316***	−0.325***						
		Rep 1	−0.090	0.284***	−0.228***						
		Rep 2	0.049	0.160***	−0.212***						
Implicit causality object-bias	Harm	*Study 1*	0.311***	0.000	0.271***	−0.097*					
		Rep 1	0.253***	−0.063	0.161*	−0.224***					
		Rep 2	0.122***	−0.074	0.200***	−0.177***					
	Filler	*Study 1*	0.082	0.001	−0.024	0.062	0.206***				
		Rep 1	0.097	0.068	0.008	−0.007	0.292***				
		Rep 2	0.081*	−0.007	0.68	−0.041	0.155***				
Explicit causal ratings	Harm: Agent	*Study 1*	−0.227***	0.164***	−0.235***	0.300***	−0.296***	0.079			
	Harm: Patient	*Study 1*	0.271***	−0.161***	0.312***	−0.251***	0.342***	−0.042	−0.515***		
	Fillers: Agent	*Study 1*	0.021	0.153***	−0.019	0.107*	−0.003	0.126**	0.162***	0.027	
	Fillers: Patient	*Study 1*	0.194***	0.024	0.166***	−0.015	0.140**	0.088	0.020	0.353***	0.310***

Study 1 (n = 459), Rep 1 (n = 249), Rep 2 (n = 788). Binding = Binding values. Individ. = Individualizing values. Contam. = Ratings of victims as contaminated. Injured = Ratings of victims as injured. Harm = Implicit causality object-bias for events of harm and force. Fillers = Implicit causality object-bias for neutral events. Harm: Agent = Causal contribution of agents for events of harm and force. Harm: Patient = Causal contribution of patients for events of harm and force. Fillers: Agent = Causal contribution of agents for neutral events. Fillers: Patient = Causal contribution of patients for neutral events. ***$p < .001$, **$p < .01$, *$p < .05$, not corrected for multiple comparisons.

"*Patient Contribution*" aggregate variable by averaging patient control, patient allowing, and patient deserving ratings.

A series of correlations indicated that, as hypothesized, contamination ratings were significantly associated with a more pronounced implicit causality object-bias, increased patient contribution ratings, and decreased agent contribution ratings. By contrast, ratings of injury were significantly negatively associated with implicit causality object-bias and with ratings of patients as causal contributors. They were also significantly positively associated with agent contribution ratings (Table 5).

People's ratings of how "contaminated" and "injured" they considered generic, unnamed victims (i.e., of rape, stabbing, strangling) was reliably related to implicit causality object-bias—that is, selecting the object as the cause for harm/force events (such as "*Bob killed Amy because. . .she"*).

Replicating prior work, and supporting our third hypothesis, binding values were positively correlated with stigmatization, and individualizing values were positively correlated with sensitivity to suffering. Furthermore, and in support of our fourth hypothesis, the implicit causality object-bias for harm and force events was associated with explicitly less sensitivity to suffering, judgments of agents as less necessary and sufficient, greater stigmatization, and increased judgments that patients allowed, controlled, and deserved harm and force events.

6. General discussion

Prior work found that blaming victims, stigmatizing them as contaminated, and viewing them as responsible for their circumstances are associated with endorsing moral values aimed at keeping groups intact: the "binding values" of loyalty, respect for authority, and purity (Graham et al., 2011; Niemi & Young, 2016). Because participants' judgments of blame in this prior work were mediated by judgments of responsibility, it was conjectured that binding values might involve a different understanding of the causal locus of events of harm and force. The results of the present research suggest that people who more strongly endorse binding values interpret the causal locus of harm events differently, as indicated (a) by their responses to the implicit causality task and (b) by their explicit causal judgments. We found that people with more strongly endorsed binding values were more likely to attribute causation to sentence objects over sentence subjects in the implicit causality task across a range of harm and force events. People high in binding values also rated sentence objects (patients) as more likely to have allowed, controlled, and deserved harm and force, and sentence subjects (agents) as less necessary and sufficient for harm and force.

One possible account of the present results is that events that violate binding values fit poorly with the "moral dyad" (Gray et al., 2012). Thus, people high in binding values are more likely to condemn violations that do not involve a dyadic structure—that is, it is more likely that binding violations can be judged as wrong without identifying an agent and a patient. When people high in binding values encounter cases of dyadic harm, they

may be more prepared to condemn the violation without drawing upon the agent-patient moral dyad template. This may contribute to their greater likelihood of assigning the causal locus for harm to affected patients, relative to people low in binding values.

The explicit causal judgment results revealed that people higher in binding values exhibit a "hydraulic" understanding of causation in a dyadic context, despite the fact that their judgments were more likely to deviate from the template in which agents are causal, and patients are not. They were more likely to judge patients as having allowed, controlled, and deserved harm. At the same time, they were less likely to judge agents of harm as necessary and sufficient. That is, the more causal they perceived patients, the less causal they perceived agents.

Prior work suggests that people who strongly endorse individualizing values also approach moral judgment in a dyadic context hydraulically: Individualizing values are associated with viewing perpetrators as making greater contributions to harm, and victims as injured by harms (Niemi & Young, 2016). In the current research, similar associations were observed: Individualizing values were related to explicit judgments of agents' causal contributions to harm, and also with increased sensitivity to the suffering of patients (Table 5). Sensitivity to suffering, but not individualizing values, was associated with selecting the subject (the harm-doer) in the implicit causality task (in two studies; Table 5). This dissociation of sensitivity to suffering and individualizing values indicates an interesting area for further research: While individualizing values have been repeatedly found to predict sensitivity to suffering, selecting the agent as causal for harm in the implicit causality task is predicted by sensitivity to suffering rather than individualizing values.

It may seem counterintuitive that binding values—*moral* values—are associated with an understanding of causation of harm that places the causal locus on a person *affected* by harm/force. The link between binding values and object-bias may be motivationally supported in at least two ways. First, the responses of people high in binding values might be driven by relatively benign motives—in spite of our finding that object-bias for harm correlated with ratings of patients as contaminated, and patients as more likely to have allowed, controlled, and deserved harmful events. We did not measure participants' concern about recklessness, negligence, prudence, or social harmony. It is possible that people higher in binding values exhibit increased concerns about *victims* as imprudent triggers of events that bring harm upon themselves, in keeping with increased concerns about group-level order and social harmony.

Alternatively, the findings could represent a motivated shift in understanding of causation (Alicke, 2000; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015) that comes along with moralized judgments that chastise underdogs to protect the status of one's own self and associates. This interpretation would be consistent with the purported function of binding values to motivate behavior and attitudes that prioritize ingroups and the family unit above the suffering of any one individual (e.g., Graham et al., 2011; Haidt, 2007). This account is also consistent with prior findings of a positive association between binding values and status-seeking (Niemi & Young, 2013).

6.1. The affordances of multiple measures of causality

Is performance on the implicit causality task implicit? One reason to believe it is implicit is the connection between rapid language production and comprehension and implicit beliefs and biases. For example, research using multiple measures of language processing, including timing the pace of reading one word at a time, reveals gender bias (von der Malsburg et al., 2020). Moreover, the task is especially lean and covert by design; task instructions are neutral as to content, and accompanying measures (e.g., for moral values) are embedded among other surveys. Nevertheless, the sources influencing participants' responses as they complete the task are not necessarily implicit. The current research demonstrates that people's moral values are reliably associated with causal attributions to people affected by harm, *whether or not* these attributions can be categorized as implicit.

Regarding the explicit measures of causality, the current work is the first to link implicit causality selections with people's judgments of agents' necessity and sufficiency, and patients' allowing, controlling, and deserving of events. Other researchers have examined how verbs' implicit causal biases vary with other kinds of causally relevant information about people (e.g., covariation; Brown & Fish, 1983; Rudolph, 2008). Our aim went beyond linking implicit causality behavior and explicit causal judgments: We examined the relationship between these measures and both individual differences in stable moral values and moral judgments of situations.

We chose to assess agents' necessity and sufficiency as these are typically considered conditions relevant to being the cause. To make sure that the way in which we ask these questions was not confounded with explicit moral judgments, we asked, "Would [Amy] have been [coerced] by someone else?" to assess necessity, and "Would [Bob] [coerce] someone else?" to assess sufficiency. One might argue that in order to answer these questions, participants still had to ask themselves whether Amy was a pushover, or whether Bob was manipulative. We could also have asked more directly: "Who caused the rape? Amy or Bob?" It's likely that such direct questioning would have alerted participants to social desirability concerns or triggered reactive affect about victim-blaming.

Measuring participants' judgments of agents' and patients' explicit causal contributions more covertly with multiple questions not only helped to circumvent social desirability concerns but also enabled participants to make their judgments more flexibly. Instead of using a bipolar scale such as "agent-caused versus patient-caused," which would require participants to treat causation in a zero-sum manner across the dyad, our items measuring agents' necessity and sufficiency and patients' capacity to control, allow, and deserve the event let us determine whether participants treated agents' and patients' contributions in a zero-sum manner even when measured in an unconstrained way. As noted above, we found that participants do indeed treat agents' and patients' explicit contributions as though they are hydraulically related (i.e., when agents are rated more causal, patients are rated less causal). In addition, explicit responses correlated with implicit causality responses, which are bipolar in nature. Ultimately, the use of multiple explicit causality items with scaled response options revealed that higher endorsement of binding values is

not just associated with broad over-attribution of causation to agents and patients of harm —binding values are associated specifically with over-attribution to *patients*.

Inquiring about participants' causal selections and explicit moral values and judgments also allowed us to observe whether and how these variables interrelated. Most notably, because of the potential consequences for harmed people, binding values of loyalty, obedience to authority, and preservation of purity were related to stigmatization of victims (replicating previous findings; Niemi & Young, 2016), increased explicit causal attributions to patients and reduced causal attributions to agents, and implicit causality object-bias for harm and force. No relationship was observed between binding values and sensitivity to victims' suffering (ratings of victims as injured). By contrast, implicit causality object-bias for harm and force correlated with reduced sensitivity for victim suffering in two studies. This finding suggests two potential sources driving object-bias for harm in the implicit causality task: (a) callousness and (b) moral (binding) values.

This research demonstrates the advantages of measures that tap multiple levels of awareness and, in particular, the advantages of the implicit causality task as a measure of people's intuitions about causation in the case of harm and force. Since the task is repeated over several trials, the experimenter can embed numerous foils, including positive and neutral events. As the response options are limited to just "he" or "she," people are likely to underestimate the extent to which any individual choice may be informative.

Ultimately, we found that stripping down a range of events involving harm and force to their most minimal possible descriptions (e.g., "*Bob coerced Amy because*") and determining the likelihood that participants select the object as referent results in an informative measure about morality. Most reliably, this approach sheds light on people's tendencies toward victim stigmatization and their moral commitments: their valuation of loyalty, obedience to authority, and concern about preservation of purity.

These latter social-moral attitudes are attitudes that those in military, legal, and clinical settings, who lead, litigate, and care for harmed people might prefer to guard or conceal. Thus, there is viable research utility for the implicit causality task, for example, in testing its use as a covert measure of attitudes toward stigmatized populations, such as sexual assault victims or minorities in various settings. More broadly, in diverse organizational settings, it is important that attributions of causation can be measured covertly with tools like the implicit causality task, as they have the potential to inform understanding of people's moral attitudes (Banaji & Heiphetz, 2010; Greenwald & Banaji, 1995; Greenwald et al., 2002; see next section).

6.2. Implicit causality task as a social psychology tool

We show how an instrument from psycholinguistics can be used to examine social-moral cognition. Our results indicate that the implicit causality task is a promising social psychology tool that reveals how selections of the causal locus for an event are shaped by social psychological factors—specifically, a measure of moral values. The implicit causality task is an efficient language measure that can be adapted for many kinds of questions relevant to social cognition. We note that this raises a challenge for the implicit

causality literature, as current theories tend to argue that non-linguistic cognition is either always relevant for implicit causality or never relevant for implicit causality. Certainly, implicit verb causality is a nuanced phenomenon (Bott & Solstad, 2014; Ferstl et al., 2011; Fiedler & Krüger, 2014; Garvey & Caramazza, 1974; Hartshorne, 2013; Hartshorne & Snedeker, 2012; LaFrance, Brownell, & Hahn, 1997; Pickering & Majid, 2007; Rudolph & Forsterling, 1997). Our aim here is not to test between theories of implicit causality per se. Instead, we propose additional research in this challenging area, as it intersects with social and moral cognition.

Some researchers have proposed that implicit causality responses may differ systematically across different sorts of verbs because people draw on their experience with typical causes of those sorts of events (Bott & Solstad, 2014). However, evidence has been inconsistent, and the largest and most systematic investigations provide limited support for this claim (e.g., Ferstl et al., 2011). Work examining the role of personal experience in causal attributions such as implicit causality responses has not been systematic. Future research should invoke responses for self and other; use sentence completions with the implicit causality task; and vary the nature of the verb. For example, personal experience with verbs conveying morally complex events (e.g., *raped, assaulted*) is likely to be associated with different results compared to neutral verbs. In the current work, individual variation in moral values was linked to implicit causality task performance specifically for morally relevant events. For these events, people higher in binding values were more likely to select the object as the cause, compared to people low in binding values. There was no consistent effect on neutral events. Likewise, personal experience may be not always be relevant to implicit causality responses, but it might have an effect if it is very salient, for example, for morally relevant events in particular: People who have been "agents" or "patients" of moral events might be less ambivalent about the causal locus for those events (i.e., whether the agent versus the patient typically causes that sort of event). However, personal experience is unlikely to fully account for the general pattern we observed involving binding values predicting increased object-bias across a range of harm/force events.

Future work will need to examine whether people higher in binding values are more likely to select the object as causal in the implicit causality task because they have a broader temporal representation of harm and force events that presupposes a prior event in which the patient performed a bad action that made them deserving of "punishment" by the agent (cf. Bott & Solstad, 2014). Here, a discourse semantics analysis of sentence completion data combined with a survey of moral values can illuminate potential differences in event representations. For example, is object-selection driven by a view of various kinds of harm/force events as "deserved punishment"? Indeed, object-bias was associated with explicit judgments of patients as deserving of harm/force events—part of the patient contribution variable. Whether these judgments reflect a broader representation of harm/force events involving an additional, triggering prior event in which the patient was causal remains to be investigated. Ongoing research bridging moral psychology and discourse semantics can shed light on these findings. Research that combines measures including sentence completion and descriptive text collection with the implicit causality

task and survey of moral values can help reveal potential differences in explanation styles, presuppositions, and event representations.

7. Conclusion

We found that a cluster of moral values—"binding values" of loyalty, obedience to authority and purity—correlated with explicit causal judgments of agents as less necessary and sufficient, and patients as more likely to have allowed, controlled, and deserved the harmful outcomes. Endorsement of binding values predicted a shift in people's expectations about who caused the harmful events: Higher binding values were related to a greater likelihood of selecting the person who was harmed as the cause. Taken together, the results indicate that people with different moral commitments differ in where they place the causal locus of harmful events, which, in turn, relates to their explicit attitudes about stigmatization and blame. These findings demonstrate that the implicit causality task from psycholinguistics is an important tool for future research on social and moral cognition.

Author Contributions

All authors contributed to the study design, data interpretation, manuscript preparation and revisions, and approved the final version of the manuscript for submission. Testing, data collection, and analysis were performed by Laura Niemi and Matthew Stanley.

Notes

1. Attention checks failure criteria were identical across studies and included choosing 1 or 2 on a Likert scale of agreement with the item "It is better to do good than bad," or 5 or 6 on the scale measuring how relevant it was to their criteria of right or wrong: "Whether or not someone was good at math" from the Moral Foundations Questionnaire (the standard "attention check items" from the Moral Foundations Questionnaire), and completion of any of four blocks of Moral Foundations Questionnaire questions in under 10 s. Any one error was sufficient for exclusion. In Study 1, of 648, 189 attention check failures; Replication Dataset 1: 135 attention check failures, 315 failed to complete the study, likely because a lengthy pilot followed the implicit causality portions. Replication Dataset 2: 284 attention check failures.
2. Data and materials available at https://github.com/lauraniemiphd/moral_ic.
3. Two additional verbs were omitted from analyses ("confused" and "punished") over concern about their neutrality, and to balance the representation of typically subject- and object-biased verbs.

4. Similar statistical modeling approaches have been implemented in psycholinguistics research (e.g., Nappa & Arnold, 2014). We did not include random slopes because many models failed to converge when random slopes were included.
5. Harms used in prior work conducted by Niemi and Young (2016).

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives in Psychological Science*, *10*, 790–812.

Arnold, J. E. (2015). Women and men have different discourse biases for pronoun interpretation. *Discourse Processes*, *52*, 77–110.

Banaji, M. R., & Heiphetz, L. (2010). Attitudes. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 353–393). New York: John Wiley & Sons.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effect models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bott, O., & Solstad, T. (2014). From verbs to discourse: A novel account of implicit causality. In B. Hemforth, B. Mertins, & C. Fabricius-Hansen (Eds.), *Psycholinguistic approaches to meaning and understanding across languages* (pp. 213–251). Cham, Switzerland: Springer.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*, 237–273.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30–37.

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods*, *43*, 124–135.

Fiedler, K., & Krüger, T. (2014). Language and attribution: Implicit causal and dispositional information contained in words. In T. M. Holtgraves (Ed.), *The Oxford handbook of language and social psychology*. New York: Oxford University Press.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*, 459–464.

Glick, P., & Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*, 491–512.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

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

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*, 613–628.

Hartshorne, J. K. (2013). What is implicit causality? *Language, Cognition and Neuroscience*, *29*, 804–824.

Hartshorne, J. K., & Snedeker, J. (2012). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, *28*, 1474–1508.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*, 65–81.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, *42*(1), 21–40.

Kipper-Schuler, K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania.

LaFrance, M., Brownell, H., & Hahn, E. (1997). Interpersonal verbs, gender, and implicit causality. *Social Psychology Quarterly*, *60*(2), 138–152.

Levin, B. (1993). *English verb classes and alternations*. Chicago, IL: Chicago University Press.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*, 464–470.

Maass, A., Ceccarelli, R., & Rudin, S. (1996). Linguistic intergroup bias: Evidence for in-group-protective motivation. *Journal of Personality and Social Psychology*, *71*(3), 512–526.

Nappa, R., & Arnold, J. E. (2014). The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension. *Cognitive Psychology*, *70*, 58–81.

Niemi, L., Roussos, G., & Young, L. (2019). Political partisanship alters the causality implicit in verb meaning. *Journal of Language and Social Psychology*, *38*(5–6), 809–819.

Niemi, L., & Young, L. (2013). Caring across boundaries versus keeping boundaries intact: Links between moral values and interpersonal orientations. *PLoS ONE*, *8*(12), e81605.

Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*, *42*, 1227–1242.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*, 780–788.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org. Accessed October 12, 2018.

Rudolph, U. (2008). Covariation, causality, and language: Developing a causal structure of the social world. *Social Psychology*, *39*, 174–181.

Rudolph, U., & Forsterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin*, *121*, 192–218.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit gender bias in linguistic descriptions for expected events: The Cases of the 2016 United States and 2017 United Kingdom elections. *Psychological Science*, *31*(2), 115–128.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Data S1. Supplementary Material includes notes on methodological differences between Study 1 and the replication datasets; and results of analyses of Implicit Causality object-bias and gender condition, demographic controls, and individualizing values.

Appendix: 30-item Moral Foundations Questionnaire (Likert-Scale Scored from 1 to 6; Graham et al., 2011)

When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? (Not at all relevant to Extremely relevant)

1. Whether or not someone suffered emotionally ***(Caring)***.
2. Whether or not some people were treated differently than others ***(Fairness)***.
3. Whether or not someone's action showed love for his or her country ***(Ingroup Loyalty)***.
4. Whether or not someone showed a lack of respect for authority ***(Authority)***.
5. Whether or not someone violated standards of purity and decency ***(Purity)***.
6. Whether or not someone was good at math ***(Attention Check)***.
7. Whether or not someone cared for someone weak or vulnerable ***(Caring)***.
8. Whether or not someone acted unfairly ***(Fairness)***.
9. Whether or not someone did something to betray his or her group ***(Ingroup Loyalty)***.
10. Whether or not someone conformed to the traditions of society ***(Authority)***.
11. Whether or not someone did something disgusting ***(Purity)***.
12. Whether or not someone was cruel ***(Caring)***.
13. Whether or not someone was denied his or her rights ***(Fairness)***.
14. Whether or not someone showed a lack of loyalty ***(Ingroup Loyalty)***.
15. Whether or not an action caused chaos or disorder ***(Authority)***.
16. Whether or not someone acted in a way that God would approve of ***(Purity)***

Please read the following sentences and indicate your agreement or disagreement:

17. Compassion for those who are suffering is the most crucial virtue. ***(Caring)***.
18. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly. ***(Fairness)***.
19. I am proud of my country's history. ***(Ingroup Loyalty)***.
20. Respect for authority is something all children need to learn. ***(Authority)***.
21. People should not do things that are disgusting, even if no one is harmed. ***(Purity)***.
22. It is better to do good than to do bad. ***(Attention Check)***.
23. One of the worst things a person could do is hurt a defenseless animal. ***(Caring)***.
24. Justice is the most important requirement for a society. ***(Fairness)***.
25. People should be loyal to their family members, even when they have done something wrong. ***(Ingroup Loyalty)***.
26. Men and women each have different roles to play in society. ***(Authority)***.
27. I would call some acts wrong on the grounds that they are unnatural. ***(Purity)***.
28. It can never be right to kill a human being. ***(Caring)***.
29. I think it is morally wrong that rich children inherit a lot of money while poor children inherit nothing. ***(Fairness)***.

30. It is more important to be a team player than to express oneself. ***(Ingroup Loyalty)***.
31. If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty. ***(Authority)***.
32. Chastity is an important and valuable virtue. ***(Purity)***.