Moral values reveal the causality implicit in verb meaning

Laura Niemi*, University of Toronto

Joshua Hartshorne, Boston College

Tobias Gerstenberg, Stanford University

Matthew Stanley, Duke University

Liane Young, Boston College

***Corresponding author**:
Laura Niemi
University of Toronto
Munk School of Global Affairs and Public Policy
315 Bloor Street West
Toronto, ON M5S 1W7
laura.niemi@utoronto.ca

Acknowledgments. We gratefully acknowledge the helpful comments and suggestions of Fiery Cushman, Felipe De Brigard, John Doris, Joshua Greene, Thomas Icard, Joshua Knobe, Jorie Koster-Hale, Laurie Paul, Jonathan Phillips, Steven Pinker, Steve Sloman, Jesse Snedeker, CUNY 2016 attendees, the Morality Lab at BC, the Events in Language Reading Groups at Harvard and MIT, and the Moral Psychology Research Group, and funding by a Sage Young Scholar Award.

Abstract

Prior work found that “binding values,” moral values that protect the group, are linked to victim-blaming. It was speculated that binding values involve an understanding of the causal structure of harmful events in which causation is placed on harmed people. The present research used the implicit causality task from psycholinguistics (e.g., "*Bob verbed Amy because*… he or she?"), explicit judgments of causal contributions, and measurement of participants’ moral values to investigate how moral values relate to interpretation of the causal structure of events. Using two verb sets and two independent replications (*N* = 459, *N* = 249, *N* = 788), we found that binding values predicted selection of the object (victim) as the cause in the implicit causality task for harmful events. Binding values also predicted explicit causal attributions. The findings indicate that moral values that support close social bonds reliably predict causal attributions to people affected by harmful events.

Keywords: Morality; Social Cognition; Causal Attribution; Implicit Causality; Psycholinguistics; Verb Semantics; Moral Psychology; Event Cognition

Moral values reveal the causality implicit in verb meaning

People often disagree about what's right or wrong. When a negative event happens, however, they tend to ask the same sorts of questions, such as: *How did this happen?*, *Who could have let it happen?*, and *Would this have happened to somebody else, regardless?* (Alicke, 2000; Alicke, Mandel, Hilton, Gerstenberg & Lagnado, 2015; Heider, 1958; Malle, Guglielmo & Monroe, 2014). These questions highlight how—moral disagreements aside—everyone engages causal cognition when exercising moral judgment, as indicated by much work in moral psychology (Alicke, 2000; Alicke et al, 2015; Cushman, 2008; Heider, 1958; Malle et al, 2014; Shaver, 1985). At what point in the evaluation of events does moral disagreement emerge?

One explanation is that moral disagreement is largely a function of differences of opinion about the specific actions that should be considered moral or immoral (e.g., premarital sex, abortion). Indeed, there are clear patterns in what people do or do not find immoral; in particular, at least two clusters of moral values – (i) caring and fairness values, deemed "individualizing values" because they are extended to each individual regardless of group membership; and (ii) loyalty, obedience, and purity values, deemed "binding values" because they purportedly keep people bound into relationships and groups (Graham et al., 2011).

These two clusters of values could lead to moral disagreement not only because of differences in what is regarded as immoral, but because they involve very different perspectives on cause and effect. Violations in the first cluster, (i) individualizing values, involve an agent causing harm to a patient (the "moral dyad"— Gray, Young & Waytz, 2012). Violations in the second cluster, (ii) binding values, may not involve any obvious harm done by an agent to a patient. Indeed, they don't require clear boundaries between the roles of agent and patient, may require a third person, and can sometimes be performed solely by an agent (Niemi & Young, 2016). Moreover, the same event (*A killed B*) may be immoral from the perspective of individualizing values but morally *obligatory* from the perspective of binding values (*A was* *ordered to kill the traitor B*). This shift in moral acceptability is made possible by viewing the killing from an alternative causal perspective in which *A*'s causal contribution as the agent is reduced and B's contribution as the patient is increased.

People tend to systematically differ in their endorsement of binding and individualizing values based on politics: binding values are higher in people who more strongly endorse conservative political ideology (Graham, Haidt & Nosek, 2009; Graham et al., 2011). Prior research has demonstrated that cultural differences (i.e., collectivist/individualist; East-West) in both value clusters are small however, a surprising finding given cultural variation in moral judgments (Feinberg, Fang, Liu & Peng, 2019) and causal cognition (e.g., Morris & Peng, 1994) that indicates collectivists value social causes more than individualists. While both value clusters are observed universally, variation is generally expected *within* cultures around binding values, and based most consistently on politics (Graham et al., 2011). Previously, Niemi and Young (2016) found increased explicit victim-blaming in people higher in binding values, even after controlling for political orientation. Ratings of victims' responsibility mediated the relationship between victim-blaming and binding values, in line with work showing that causal judgments feed judgments of blame (e.g., Malle et al., 2012). It was speculated that binding values might involve altered *representation of the causal structure of harm*, leading people higher in binding values to be more likely to shift causation off of harm-doers and over to harmed people. The present research directly investigates this possibility.

We first investigate this hypothesis using a task that measures intuitions about likely causes for events from psycholinguistics, the implicit causality (IC) task (Brown & Fish, 1983; Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Rudolph & Forsterling, 1997). In this task, participants read sentences such as

(a) Bob murdered Amy because…

and chose whether to continue the sentence with a pronoun referring to the agent ("he") or patient ("she"). Continuing (a) with *he* or *she* reveals an expectation that the murder is due to something that the subject (Bob) did or the object (Amy) did, respectively.

Implicit causality research indicates that this choice tends to vary by verb (*praise*, *frighten*, etc.). Some verbs reliably prompt selection of the pronoun that refers to the subject (e.g., “subject biased” verbs like *frighten*); some prompt selection of the pronoun that refers to the object (e.g., “object biased” verbs like *praise*), suggesting that people have systematic expectations about how some categories of events came about (Brown & Fish, 1983; Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Rudolph & Forsterling, 1997; Bott & Solstad, 2014; Ferstl, Garnham & Manouilidou, 2011; Hartshorne, 2013; Pickering & Majid, 2007; Rudolph, 2008). At this point, the only well-established predictor of IC is verb semantics; specifically, when verbs have been clustered into classes based on fine-grained analyses of shared semantic and syntactic features — which include information about causation — they tend to share implicit causality biases to the subject or object (Hartshorne, 2013; Hartshorne & Snedeker, 2012; Kipper-Schuler, 2006).

Interestingly, although many researchers have suggested that IC is broadly affected by individuals’ beliefs about the world, such as perceived social hierarchy or gender roles, the accompanying evidence has been inconsistent (Garvey & Caramazza, 1974; Bott & Solstad, 2014; Ferstl, Garnham & Manouilidou, 2011; Hartshorne, 2013; Pickering & Majid, 2007). Despite the potential for IC to be affected by relevant individual differences (e.g., one’s moral values as they consider morally relevant events, this topic is still relatively unexplored – with the exception of work examining gender differences in pronoun comprehension and interpretation (Arnold, 2015). Social psychology of language researchers have traced other lexical features through which people convey and modulate moral judgments, largely focusing on character judgments. For example, the higher moral stakes of using abstract adjectives *versus* action verbs to describe a person’s contribution to an event (Fiedler & Krüger, 2014) and the implications of withholding negative adjective labels to protect the ingroup (linguistic ingroup bias, e.g., Maass, Ceccarelli, Rudin, 1996) — are well-described. The current work connects to this prior research by examining how people's stable moral values relate to their general understanding of the causal structure of moral events.

To engage in smooth dialogue, people share an understanding about how an event was caused; when their explanations don't match up, conflict can unfold (Pickering & Garrod, 2014; Taylor, 2014). IC selections provide a window into how people think about the causes of events, since explanations typically point to the most relevant or contributory causes (e.g., Hilton, 1990; Hesslow, 1988; Lombrozo, 2006). For instance, note that explaining (a) in terms of Bob is more consistent with the moral dyad framework, in which active perpetrators harm passive victims (Gray, Young & Waytz, 2012). The present research leveraged the IC task to investigate the causal structure consistent with the explicit causal and moral judgments of people ranging in moral values, and also let us contribute to research on individual differences in IC responses.

We examined whether people high in binding values were indeed more likely to exhibit an object-bias in the IC task, given prior repeatedly replicated findings that binding values predict victim blame and stigmatization (Niemi & Young, 2016). Less consistent prior evidence linking individualizing values and perpetrator blame suggested that individualizing values might be linked to the opposing pattern, subject-bias for harm events. We also tested responses to morally irrelevant, neutral verbs to rule out the possibility that people high in binding values were generally more likely to consider affected people to be causal contributors.

We also 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 we expected individuals who prioritize binding values to be less likely to apply the moral dyad framework in which agents are causal and blameworthy ("agent-harmed-patient") when reasoning about immoral events, we expected these participants to view agents as less necessary or sufficient and to view patients as more likely to have allowed, controlled or deserved the events. Moreover, we expected that IC object-bias would be directly related to judgments that agents were less necessary and sufficient and that patients allowed, controlled, and deserved the events.

We measured participants' sensitivity to victim suffering (how "injured" participants considered victims) and stigmatization of victims (how "contaminated" participants considered victims) to understand how these explicit morally-relevant attitudes about harmed people (sensitivity vs. stigmatization) related to implicit causality selections, and to test replication of prior work. In prior work (Niemi & Young, 2016), increased sensitivity to victim suffering – rating victims as more "injured" – was associated with higher individualizing values. Increased stigmatization of victims – rating victims as more "contaminated", like victim-blaming, was associated with higher binding values. Victim injury ratings were also negatively correlated with victim contamination ratings, suggesting that viewing victims as contaminated is inconsistent with viewing them as passive victims of harm. Moral values were measured with the Moral Foundations Questionnaire, which has been used extensively in prior work to measure people's diverse moral values (e.g., Graham et al., 2011; Niemi & Young, 2016).

Roadmap

We outline the four tested hypotheses in the present research below. The first three examine how moral values are related to IC responses, explicit causal judgments, and the propensity to stigmatize (or be sensitive to) victims. The last hypothesis examines their interrelationships.

(1) In the IC task, people higher in binding values were expected to 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 were expected to be negatively related to explicit causal judgments of the agent's necessity and sufficiency, and positively related to judgments of the patient's capacity to allow, control and deserve events.

(3) Positive correlations were expected between binding values and stigmatization of victims; and between individualizing values and sensitivity to victim suffering.

(4) IC object-bias for harm and force events was expected to be related to reduced sensitivity to victim suffering and judgments of agents as less necessary and sufficient, as well as greater stigmatization of victims and increased judgments that patients allowed, controlled and deserved harm and force events.

Study 1 begins by testing all four hypotheses. We attempt to replicate the findings involving the IC bias in Study 1 in Replication Dataset 1. We attempt to replicate the findings involving the IC bias using an expanded set of verbs and a larger sample size in Replication Dataset 2.

Study 1

Method

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 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 collected data from additional participants online[2] via Amazon's Mechanical Turk and attempted 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 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 missing). Data and materials are available at https://github.com/BLINDEDFORREVIEW/. Methodological differences between Study 1 and the Replication Datasets are described in the **Supplementary Materials**. The institutional review board at Boston College approved all studies, and informed consent was obtained via an online form from all participants.

Moral values in the five foundations (caring, fairness, loyalty, obedience to authority, and purity) were assessed using the 30-item Moral Foundations Questionnaire (MFQ; Graham et al, 2011). 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 ASI).

The implicit causality task involved 24 minimal event descriptions in the form: "[*Subject*] *verb-ed* [*Object*] *because*..." – e.g., *"Bob coerced Amy because..."* – with half of the participants receiving male sentence subjects and female sentence objects, and vice versa for the other half in order 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), and 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:*	.39	.41	.38
clobbered	.55	.57	.66
coerced	.30	.36	.27
enslaved	.36	.40	.24
forced	.39	.37	.34
influenced	.31	.31	.26
killed	.53	.57	.50
manipulated	.29	.32	.25
raped	.30	.31	.19
robbed	.26	.28	.31
stabbed	.50	.54	.41
strangled	.54	.55	.44
tempted	.31	.33	.26
assaulted			.41
convinced			.35
enticed			.30
groped			.32
impaled			.65
molested			.22
persuaded			.34
pressured			.37
seduced			.27
silenced			.68
spanked			.66
walloped			.66
Neutral Fillers *Mean:*	.61	.61	.58
approached	.39	.41	.35
congratulated	.89	.88	.91
delighted	.33	.34	.34
impressed	.21	.25	.21
observed	.60	.64	.58
praised	.85	.85	.89
quoted	.65	.61	.68
skipped	.57	.59	.54
thanked	.85	.86	.86
transported	.77	.71	.81
boggled			.36
caressed			.40
celebrated			.84
comforted			.82
appraised			.65
complimented			.82
honored			.86
massaged			.61
diverted			.49
fondled			.26
greeted			.51
puzzled			.23
tickled			.52
raced			.33

Note. Higher value indicates greater object-bias: in the implicit causality task, selection of the referent to the object coded as 1, subject as 0.

To assess explicit causal judgments, we gathered participants' beliefs about agents' and patients' causal contributions. 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:

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

Participants responded using sliding scales (0 = "*Definitely No*", 50 = "*Unsure*", 100 = "*Definitely Yes*").

Finally, we measured sensitivity versus stigmatization toward victims 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 prior work, only these four crimes were used to obtain measures of how "*contaminated / tainted*" and "*injured / wounded*" participants rated hypothetical crime victims. Average ratings of victims across events as contaminated/tainted and injured/wounded served as indices of stigmatization of victims and sensitivity to victim suffering, respectively.

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 victim suffering, we computed a series of Spearman's rank-order correlations.

Results and Discussion

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

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 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 Replication Datasets (see **Table 2**). To further interrogate these significant interaction effects, generalized linear mixed-effects models were computed for harm/force verbs and neutral filler verbs, taken separately.

Table 2. *The results of two 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**
Study 1					
Verb Type	2.17	.43	5.06	< .0001	[1.33, 3.02]
Binding Values	.37	.05	7.02	< .0001	[.27, .48]
Verb Type x Binding Values	-.27	.05	-5.60	< .0001	[-.36, -.18]
Replication Dataset 1					
Verb Type	2.43	.47	5.17	< .0001	[1.51, 3.36]
Binding Values	.43	.08	5.27	< .0001	[.27, .59]
Verb Type x Binding Values	-.35	.07	-4.76	< .0001	[-.49, -.20]
Replication Dataset 2					
Verb Type	1.28	.32	4.35	< .0001	[.75, 2.00]
Binding Values	.19	.04	4.66	< .0001	[.11, .27]
Verb Type x Binding Values	-.10	.03	-2.54	.0004	[-.15, -.04]

Note. Study 1 (*N* = 459), Replication Dataset 1 (*N* = 249), Replication Dataset 2 (*N* = 788). All 95% CIs are for the beta-estimates.

For harm/force verbs only, a generalized linear mixed-effects regression model was computed for which binding values was 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 effect of binding values on the likelihood of selecting the object as the referent (*p* < .0001). We obtained the same pattern of results in Replication Dataset 1 (*p* < .0001) and Replication Dataset 2 (*p* < .0001). 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.

Importantly, for the neutral filler verbs, there was no significant effect of binding values on selection of the object *relative to* the subject as the referent in the primary study or Replication Dataset 1 (both *p*s > .05). For the neutral filler verbs in Replication Dataset 2, there was a small but significant effect of binding values on the selection of the object *relative to* the subject as the referent (*p* = .02). So, the effect was much larger for harm/force verbs than for neutral filler verbs.

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. Analyses showed that binding values remain consistent significant predictors of implicit causality object-bias for harm/force verbs after controlling for political orientation, gender, and religiosity in the three datasets. Second, we found no effect of individualizing values on 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 *versus* female-verbed-male) was related to the implicit causality object-bias for harm/force verbs. Specifically, in all datasets, participants were more likely to select men for harm/force events. However, binding values continued to significantly predict the implicit causality object-bias despite the gender effect.

Overall, the implicit causality results support our first hypothesis. Participants higher in binding values were more likely to select the object over the subject as the referent for harm and force events, but not for neutral events. These effects held after controlling for a variety of other variables.

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; as a comparison, we investigated whether the opposite patterns would be observed in the case of individualizing values. We first computed a series of linear mixed-effects models in which binding 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 all five models, there was a significant interaction effect between binding values and verb type (see **Table 3**). To further interrogate these significant interaction effects, linear mixed-effects models were computed for harm/force verbs and neutral filler verbs, taken separately.

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	.78	-4.33	< .0001	[-4.75, -1.86]
Verb Type x Binding Values	2.96	.40	7.37	< .0001	[2.21, 3.74]
Outcome: Sufficiency					
Verb Type	-3.32	1.87	-1.70	.095	[-7.13, .62]
Binding Values	-1.77	.83	-2.14	.033	[-3.55, -.16]
Verb Type x Binding Values	1.18	.35	3.40	.0007	[.44, 1.87]
Outcome: Allow					
Verb Type	30.95	6.51	4.76	< .0001	[18.59, 44.02]
Binding Values	5.71	.77	7.42	< .0001	[4.16, 7.37]
Verb Type x Binding Values	-1.71	.49	-3.52	< .0001	[-2.61, -.71]
Outcome: Control					
Verb Type	18.30	5.90	3.10	.005	[7.89, 30.81]
Binding Values	4.08	.74	5.54	< .0001	[2.70, 5.54]
Verb Type x Binding Values	-1.22	.48	-2.53	.012	[-2.22, -.39]
Outcome: Deserve					
Verb Type	47.38	5.58	8.49	< .0001	[35.69, 57.98]
Binding Values	3.83	.66	5.80	< .0001	[2.39, 5.11]
Verb Type x Binding Values	-3.23	.45	-7.17	< .0001	[-4.04, -2.34]

Note. All 95% CIs are for the beta-estimates.

For harm/force verbs only, five linear mixed-effects models with binding values as the fixed predictor of necessity, sufficiency, allowing, controlling, or deserving judgments (in separate models) revealed that binding values were negatively related to participants' judgments about the agent's necessity (*p* = .0002) and sufficiency (*p* = .046), and positively related to their judgments about the patient's capacity to allow (*p* < .0001), control (*p* < .0001), and deserve (*p* < .0001) the events. Importantly, for the neutral filler verbs, there was no significant effect of binding values on judgments of necessity, sufficiency, or desert (all *p*s >

.05). However, there were significant effects of binding values on judgments of allowing (p < .0001) and controlling (p = .0006). Nevertheless, for allowing and controlling judgments, the magnitude of the effect was larger for harm/force verbs than for neutral filler verbs.

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 necessity, allowing, controlling, and deserving, there were significant interaction effects between individualizing values and verb type (see **Table 4**). 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 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, or deserving judgments (in separate models) revealed that individualizing values were not significantly related to participants' judgments about the agent's necessity (p = .88), but individualizing values were negatively related to judgments of the patient's capacity to allow (p = .048), control (p = .0006), and deserve (p = .001) the events. For the neutral filler verbs, there was only a significant effect of individualizing values on judgments of deserving (p = .016). There were no significant effects for judgments of necessity, allowing, or controlling for the neutral filler verbs (all ps > .05).

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: they were negatively related to judgments about the patient's capacity to allow, control, and deserve the events.

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	.20	1.14	.18	.859	[-1.93, 2.39]
Verb Type x Individualizing Values	-2.15	.59	-3.66	.0003	[-3.30, -.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 x Individualizing Values	-.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, -.40]
Verb Type x Individualizing Values	2.66	.71	3.75	.0002	[1.42, 3.89]
Outcome: Control					
Verb Type	4.68	6.55	.71	.480	[-8.51, 18.00]
Individualizing Values	-4.14	1.09	-3.78	.0002	[-6.21, -2.09]
Verb Type x Individualizing Values	1.91	.71	2.70	.007	[.56, 3.19]
Outcome: Deserve					
Verb Type	3.68	6.17	.60	.555	[-9.50, 15.34]
Individualizing Values	-4.16	.98	-4.24	< .0001	[-6.11, -2.27]
Verb Type x Individualizing Values	6.65	.66	10.12	< .0001	[5.39, 7.89]

Note. All 95% CIs are for the beta-estimates.

Sensitivity versus Stigmatization toward Victims. First, we computed a series of correlations to test replication of previously observed relationships between binding values and stigmatization of victims, and individualizing values and sensitivity to victim 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 victims as contaminated, and a positive relationship between individualizing values and ratings of victims as injured (see **Table 5**).

Next, to address the subsequent hypotheses that stigmatization and sensitivity for victims might be related to implicit causality object-bias and 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 (akin to an “implicit victim-blaming” score). Correspondingly, the “neutral filler object-bias” represented a tendency to select the object over the subject across events that do not involve harm and force. Additionally, we created an “*Agent Contribution*” aggregate variable by averaging the agent unnecessary ratings (reverse-coded) and agent sufficiency ratings, and a “*Patient Contribution*” aggregate variable by averaging patient control, patient allowing, and patient deserving ratings.

A series of correlations indicated that, as hypothesized, ratings of victims as contaminated were significantly associated with a more pronounced implicit causality object-bias, increased patient contribution ratings, and decreased agent contribution ratings. By contrast, ratings of victims as injured 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**). It is notable that people’s ratings of how “contaminated” and “injured” they considered generic, unnamed victims of crimes (i.e., rape, stabbing, strangling) – completed in a separate part of the study – showed reliable relationships with implicit causality object-bias — i.e., selection of the object as the cause for various harm and force events (such as: “Bob killed Amy because…*she”*).

Overall, these results largely supported our third and fourth hypotheses. Replicating prior work, binding values were positively correlated with stigmatization of victims, and individualizing values were positively correlated with sensitivity to victim suffering. The implicit causality object-bias for harm and force events was associated with explicitly less sensitivity to victim suffering, judgments of agents as less necessary and sufficient, greater explicit stigmatization of victims, and increased judgments that patients allowed, controlled and deserved harm and force events.

Table 5. *Spearman's rank-order correlations among moral values, judgments of victims as contaminated and injured, implicit causality object-bias for harm and filler events, and explicit causal judgments for agents and patients of harm/force and neutral filler events.*

							IC object-bias		Explicit causal ratings			
			Binding	Individ.	Contam.	Injured	Harm	Filler	Harm: Agent	Harm: Patient	Fillers: Agent	Fillers: Patient
Individ.		*Study 1*	.047									
		Rep 1	.197									
		Rep 2	.139**									
Contam.		*Study 1*	.473***	-.101								
		Rep 1	.402***	-.086								
		Rep 2	.368***	.053								
Injured		*Study 1*	-.183***	.316***	-.325***							
		Rep 1	-.090	.284***	-.228***							
		Rep 2	.049	.160***	-.212***							
IC Object-Bias	**Harm**	*Study 1*	.311***	.000	.271***	-.097						
		Rep 1	.253***	-.063	.161	-.224***						
		Rep 2	.147***	-.080	.201***	-.161***						
	Filler	*Study 1*	.082	.001	-.024	.062	.206***					
		Rep 1	.097	.068	.008	-.007	.292***					
		Rep 2	.111	-.021	.088	-.036	.364***					
Explicit causal ratings	**Harm: Agent**	*Study 1*	-.227***	.164***	-.235***	.300***	-.296***	.079				
	Harm: Patient	*Study 1*	.271***	-.161	.312***	-.251***	.342***	-.042	-.515***			
	Fillers: Agent	*Study 1*	.021	.153	-.019	.107*	-.003	.126	.162	.027		
	Fillers: Patient	*Study 1*	.194***	.024	.166***	-.015	.140	.155	.020	.353***	.310***	

Note. Study 1 (*N* = 459), Replication 1 (*N* = 249), Replication 2 (*N* = 788). Binding = Binding values. Individ. = Individualizing values. Contam. = Ratings of victims as contaminated. Injured = Ratings of victims as injured. Harm = IC object-bias: Implicit causality object-bias for events of harm and force. Fillers = IC object-bias: Implicit causality object-bias for neutral filler 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.

General Discussion

*Moral Values and Causal Attribution***.** Prior work found that moral values aimed at protecting groups and keeping people bound into tight-knit relationships – the "binding values" of loyalty, obedience to authority and purity (e.g., Graham et al., 2011) are associated with victim-blaming, stigmatizing judgments of victims as contaminated and tainted, and viewing victims as responsible (Niemi & Young, 2016). Because blame allocations were mediated by judgments of victim responsibility, it was speculated that binding values might involve a different understanding of the causal structure of events of harm and force. The results of the present research suggest that people who highly endorse binding values might interpret the causal structure of events slightly differently, as indicated by their responses to the implicit causality task, and by their explicit causal judgments. People who highly endorsed binding values are more likely to attribute causation to sentence objects over subjects in the implicit causality task across a range of harm and force events. In line with the interpretation that people high in binding values might interpret the causal structure of these events (e.g., *"Bob coerced Amy")* differently, they 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.

Since immoral events in the case of binding values do not necessarily fit with the "moral dyad" in which an agent harmed a patient and the agent is causal whereas the affected patient is not (Gray, Young, Waytz, 2012), one possibility is that greater endorsement of binding values may lead people to spread causal responsibility across event participants – including surprising targets: victims (Niemi & Young, 2016). Another possibility that could be tested in future work is that people higher in binding values throw out the moral dyad altogether and abandon ascribing causation in the typical zero-sum, hydraulic manner (i.e., the more causal the victim is perceived, the less the perpetrator). Results from the measures of *explicit* causal judgments indicate however that people higher in binding values retain a zero-sum understanding of blame ascription. The more strongly people endorsed binding values, the more likely they were to judge patients as having allowed, controlled and deserved harm, and the less likely they were to judge agents of harm as necessary and sufficient. The more causal contribution people high in binding values perceived in victims, the less they perceived in perpetrators.

By contrast, previous work found small associations between individualizing values and increased judgments of perpetrators as causal contributors, and increased sensitivity to victims' suffering (Niemi & Young, 2016). In the current research, similar associations were observed:

individualizing values correlated with explicit judgments of agents' contributions to harm and increased sensitivity to victims' suffering (Table 5). Individualizing values were not associated with selecting the subject (the harm-doer) in the implicit causality task. However, sensitivity to victim suffering *was* related to selection of the harm-doer in the implicit causality task in two studies (Table 5). This dissociation indicates an interesting area for further research: individualizing values corresponded with sensitivity to suffering, yet only sensitivity to suffering and not individualizing values predicted object-bias for harms in the IC.

It may be counterintuitive that binding values – *moral* values – motivate an understanding of the causal structure of harm that places the causal source of the explanation on the *affected* person. The link between binding values and object-bias may come about in at least two ways. First, people high in binding values might be driven by relatively benign motives – even though object-bias for harm correlated with rating victims as contaminated and patients as more likely to have allowed, controlled, and deserved harmful events. We did not look at participants' concern about recklessness, negligence, prudence or social harmony. It is possible that people higher in binding values have increased concerns about victims (as imprudent triggers of events that bring them harm) in keeping with their increased concern about group-level order and other social moral concerns. Alternatively, the findings could represent a shift in causal structure that comes along with moralized judgments that chastise a victim in order to protect the status of the self and associates. This would be consistent with the purported function of binding values to motivate behavior and attitudes that protect groups and the family unit before being concerned about the suffering of any one individual (e.g., Graham et al., 2011; Haidt, 2007), as well as prior findings of a positive association between binding values and status-seeking (Niemi & Young, 2013).

*Implicit Causality Task as a Social Psychology Tool***.** The results indicate that the IC task from psycholinguistics is a promising social psychology tool that reveals how the causal structure of an event is shaped by social psychological factors – indexed presently by a measure of moral values. The IC 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 IC literature, as current theories tend to argue that non-linguistic cognition is either always relevant for IC or that it never is. Certainly, implicit verb causality is a nuanced phenomenon (Bott & Solstad, 2014; Ferstl et al., 2011; Fiedler & Krüger, 2014; Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Hartshorne, 2013; LaFrance, Brownell, Hahn, 1997; Pickering & Majid, 2007; Rudolph & Forsterling, 1997). Our focus here is not theories of IC *per se,* and as specified next, more research is needed in this challenging area.

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, causal structure, and IC responses has not been systematic. Future research should invoke responses for self and other; use sentence completions with the IC 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 IC 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 relevant to IC responses for morally relevant events: people who have been "moral patients" of moral events might be less ambivalent about their causal structure (*i.e.,* whether the agent versus the patient typically causes that sort of event).

Future work will also 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 survey of moral values can help shed light on potential differences in event representations. We note as well that the finding that the events’ subject/object gender mattered to implicit causality selections (i.e., people expected events involving women harming men to be better explained with reference to the object (men) compared to when men harm women) indicates that quick inferences about violence are jointly influenced by *both* stable moral values as well as gendered intuitions about deservingness for harm.

The IC task is psychologically informative based on which types of verbs are slotted into its simple format. Namely, we were able to learn about whether people would consider victims of crimes "contaminated" or “injured” and endorse “binding values”—moral values of loyalty, obedience to authority and preservation of purity—previously found to be associated with victim blaming (Niemi & Young, 2016), by observing their selections for harm and force verbs in the IC tasks – in particular their tendency to selection the referent to the object. The IC task applied to verbs with limited moral relevance, i.e., the neutral verbs, did not predict people’s moral values and judgments of victims attitudes. In all likelihood, verb semantics drove participants’ IC responses in those cases, as prior work has demonstrated that linguistic structure explains implicit causality biases when participants’ individual differences aren’t taken into account (Hartshorne, 2013; Hartshorne & Snedeker, 2012). Thus, the IC task supplemented in these ways is a useful social psychology tool; performance was distinct from performance in explicit moral measures (i.e., moral values, ratings of victims), and explicit causal judgments of agents as necessary and sufficient, and patients as having allowed, controlled or deserved the events.

As shown in Table 1, on average, the verbs did not compel strong causal inferences toward the subject or the object, although the neutral verbs were slightly more object biased relative to the harm and force verbs. At the individual verb level, particular neutral and harm/force verbs were more likely to compel inferences that the sentence subject caused the event, such as the verbs *raped, approached,* and *delighted.* Others, by contrast, reliably and strongly compelled inferences that the sentence object caused the event, such as the verbs *praised* and *congratulated.* In the case of the neutral verbs, this is consistent with known subject- and object-biases for certain verbs classes, as previously discussed (e.g., the *amuse* and *judgment* verb classes, respectively, Kipper-Schuler, 2006). Some accounts have shown that subject-biases (e.g., for *amused)* are related to explanations involving specific actions of the agents, whereas object-biases (e.g., for *congratulated)* involve presuppositions that a prior event occurred in which the object of the current event did something that was the causal impetus for the current event (Bott & Solstad, 2014). In the case of many harmful events, which don't compel a strong subject- or object-bias on the aggregate (e.g., *killed*), it is notable that people's responses reliably reflected their binding values. 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 IC task and survey of moral values, can reveal potential differences in explanation styles, presuppositions, and causal models that might underlie the present results.

*The Affordances of Multiple Measures of Causality***.** 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 desert 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. Nevertheless, we chose to assess agents' necessity and sufficiency because 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 circumvent social desirability concerns. It also enabled participants to assign their judgments more freely. Instead of using a bipolar scale such as “agent-caused vs. 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 these were measured in an unconstrained way. 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, their explicit responses correlated with implicit causality responses, which are bipolar in nature. Finally, 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 causal responsibility to agents *and* patients of harm – binding values are associated specifically with over-attribution to patients.

Inquiring about participants’ explicit moral judgments, implicit 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 judgments of patients and reduced causal judgment of 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), whereas implicit causality object-bias for harm and force *did* correlate 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: (1) callousness, and (2) moral (binding) values.

This research demonstrates the advantages of measures that tap multiples 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 a range of events involving harm and force down 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, it informs about 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, e.g., sexual assault victims, minorities in various settings. More broadly, in many organizational settings, it is important that attributions of causation can be measured covertly in the service of understanding how people think about blame and responsibility.

Conclusion. The current work indicates a mechanism in causal cognition for prior findings in which moral values were associated with judgments of patients as blameworthy. A cluster of moral values – “binding values” of loyalty, obedience to authority and purity – were directly tied to explicit causal judgments of agents as less necessary and sufficient and patients as more likely to have allowed, controlled and deserved the harmful outcomes. The implicit causality (IC) task results showed that 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 (the sentence object) as the cause. Taken together, the results indicate that people with different moral commitments might differ in how they represent the causal structure of harmful events, which in turn, relates to their explicit attitudes about stigmatization and blame.

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 BLINDED and BLINDED.

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.

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.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353-380.

Feinberg, M., Fang, R., Liu, S., & Peng, K. (2019). A World of blame to go around: Cross-cultural determinants of responsibility and punishment judgments. *Personality and Social Psychology Bulletin, 45*(4), 634–651.

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, NY: 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.

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

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.

Hartshorne, J. K., Sudo, Y., Uruwasi, M. (2013). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology, 60*, 179-196.

Heider, F. (1958). *The psychology of interpersonal relationships.* New York: John Wiley & Sons, Inc.

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, England: Harvester Press.

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

Holtgraves, T. M. (2002). *Language as social action: Social psychology and language use.* New Jersey: Lawrence Erlbaum Associates.

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

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

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

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33,* 273-286.

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.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*, 147-186.

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. Journal of Personality and Social Psychology, 67, 949-971. doi:10.1037/0022-3514.67.6.949

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. & 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.

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

Pickering, M. J. & Garrod, S. (2014). Interactive alignment and language use. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology.* New York, NY: Oxford University Press.

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

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.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.

Taylor, P. L. (2014). The role of language in conflict and conflict resolution. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology*. New York, NY: Oxford University Press.

Footnotes

[1] Attention checks failure involved responding with 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 MFQ), and completion of any of four blocks of MFQ questions in under 10 s. Original sample = 648, any one error was sufficient for exclusion. We excluded 189 additional individuals who failed attention checks.

[2] Attention checks failure criteria for Replication Datasets 1 and 2 were identical to Study 1. Replication Dataset 1 attention check failures = 135; exclusions for failing to complete the study = 315, likely because following the IC portions, this study involved a lengthy pilot. Replication Dataset 2 attention check failures = 284.

[3] Two additional verbs, "confused" and "punished," were omitted from analyses over concern about the neutrality of *punished,* and to balance the representation of typically subject- and object-biased verbs with its removal.

[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).

Supplementary Materials
Moral values reveal the causality implicit in verb meaning

A: Notes on Methodological Differences between Study 1 and the Replication Datasets

Replication Dataset 1. As in Study 1, participants completed a block of the implicit causality task and then entered a separate block; this time the events were re-presented with the pronoun they had selected in the first block (e.g., “Max verbed Jess because he…) and an empty text box appeared after the pronoun with the prompt: “please finish the sentence.” Participants typed a completion to each sentence. They also filled out measures of demographics, victim stigmatization and sensitivity, and moral values as in the previous study (data and materials available at https://github.com/ BLINDED FORREVIEW).

Replication Dataset 2. As in the previous studies, participants completed the implicit causality task, but the task here involved a total of 48 verbs. See Table 1 in the main text for the complete list of verbs. Participants also filled out measures of demographics, victim stigmatization and sensitivity, and moral values as in the previous studies. They completed the Ambivalent Sexism Inventory (Glick & Fiske, 1996); the present analyses did not involve the ASI. Data and materials are available at https://github.com/ BLINDED FORREVIEW.

B: Gender Condition and Implicit Causality Object-Bias from Study 1 and the Replication Datasets

To investigate whether the gender condition (male-verbed-female *versus* female-verbed-male) predicted the implicit causality object-bias for harm/force verbs, we first computed a generalized linear mixed-effects regression model in which verb type (harm/force (coded as 0) *versus* neutral filler (coded as 1)) and gender condition (male-verbed-female (coded as 0) *versus* female-verbed-male (coded as 1)) 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 (participant and verb were both included as random effects with random intercepts only). There was a significant interaction between verb type and gender condition in Study 1 and in the Replication Datasets (see **Supplementary Table 1**). To further interrogate these significant interaction effects, generalized linear mixed-effects models were computed for harm/force verbs and neutral filler verbs, taken separately.

Supplementary Table 1. *The results of two generalized linear mixed-effects regression models—each with verb type and gender condition as predictors of selecting the object versus the subject as the referent.*

	b	*SE*	*Z*	*p*	95% CI
Study 1					
Verb Type	1.67	.40	4.19	< .0001	[.89, 2.45]
Gender Condition	.53	.11	4.89	< .0001	[.32, .75]
Verb Type x Gender Condition	-.99	.10	-10.29	< .0001	[-1.18, -.80]
Replication Dataset 1					
Verb Type	1.52	.29	3.95	< .0001	[.77, 2.28]
Gender Condition	.71	.15	4.77	< .0001	[.42, 1.00]
Verb Type x Gender Condition	-.78	.13	-5.99	< .0001	[-1.03, -.52]
Replication Dataset 2					
Verb Type	1.45	.30	4.81	< .0001	[.86, 2.04]
Gender Condition	.62	.08	8.28	< .0001	[.48, .77]
Verb Type x Gender Condition	-.87	.05	-.17.74	< .0001	[-.97, -.78]

Note. Main Experiment (*N* = 459), Replication Dataset (*N* = 249), Replication Dataset 2 (*N* = 788). All 95% CIs are for the beta-estimates.

For harm/force verbs, a generalized linear mixed-effects regression model was computed for which gender condition was 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 (participant and verb were both included as random effects with random intercepts only). This analysis yielded a significant effect of gender condition on the likelihood of selecting the object versus the subject as the referent in Study 1 (*b* = .63, *SE* = .13, *Z* = 4.98, *p* < .0001, 95% CI = [.38, .87]). We obtained the same pattern of results in Replication Dataset 1 (*b* = .80, *SE* = .18, *Z* = 4.35, *p* < .0001, 95% CI = [.44, 1.16]) and Replication Dataset 2 (*b* = .69, *SE* = .08, *Z* = 8.31, *p* = .0001, 95% CI = [.53, .85]). In all datasets, when women harmed/forced men, participants were more likely to select the object than the subject as the referent (i.e., there was a more pronounced object-bias).

For neutral filler verbs, there was a significant effect of gender condition on the propensity for selecting the object over the subject as the referent in the main experiment (*b* = -.43, *SE* = .11, *Z* = -3.80, *p* = .0001, 95% CI = [-.66, -.21]); there was no effect in Replication Dataset 1 (*b* = -.04, *SE* = .14, *Z* = -.28, *p* = .78, 95% CI = [-.32, .24]); the effect returned in Replication Dataset 2 (*b* = -.24, *SE* = .07, *Z* = -3.47, *p* = .0005, 95% CI = [-.38, -.11]). The significant effects for neutral filler verbs were in the opposite direction of the effects for harm/force verbs. To sum up, moral values aside, participants generally were more likely to select men for harm/force events in the implicit causality task regardless of whether they were the subject (the “perpetrator” of harm/force) or the object (the “victim” of harm/force).

However, binding values withstood this gender effect. Importantly, when gender condition was added to the generalized linear mixed-effects models that already included binding values as a predictor of the propensity to select the object relative to the subject as the referent, participants higher in binding values remained significantly more likely to select the object over the subject as referent for harm/force verbs in Study 1 (*b* = .41, *SE* = .06, *Z* = 6.76, *p* < .0001, 95% CI = [.29, .52]), Replication Study 1 (*b* = .50, *SE* = .10, *Z* = 4.95, *p* < .0001, 95% CI = [.30, .69]), and Replication Study 2 (*b* = .21, *SE* = .05, *Z* = 4.65, *p* < .0001, 95% CI = [.12, .30]).

C: Demographic Controls and Implicit Causality Object-Bias

Having found that for harm/force verbs, binding values significantly predict the likelihood of selecting the object over the subject as referent ("object-bias"), we next expanded the generalized linear mixed-effects regression models to include political orientation, gender, and religiosity as additional fixed predictors along with binding values. 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. More specifically, binding values, political orientation, gender (0 = male, 1 = female), and religiosity 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 for the harm/force verbs. Full results for these models for Study 1 and the replication datasets are depicted in **Supplementary Table 2**. Binding values remained a consistent, significant predictor of the object-bias in all three datasets after statistically controlling for political orientation, gender, and religiosity. An inconsistent effect of gender appeared in two cases in Study 1 and Replication Dataset 2 such that male participants were more likely to exhibit object-bias than female participants.

Supplementary Table 2. *The results of two generalized linear mixed-effects regression models—each with binding values, political orientation, gender, and religiosity as predictors of the propensity to select the object as the referent for harm and force events.*

	b	*SE*	*Z*	*p*	95% CI
Study 1					
Binding Values	.21	.08	2.73	.006	[.06, .36]
Political Orientation	-.07	.04	-1.59	.111	[-.15, .01]
Gender	-.58	.12	-4.92	< .0001	[-.82, -.35]
Religiosity	.09	.03	2.69	.007	[.02, .15]

Replication Dataset 1					
Binding Values	.53	.13	4.08	< .0001	[.28, .79]
Political Orientation	.08	.06	1.31	.190	[-.04, .21]
Gender	-.35	.18	-1.92	.055	[-.71, .01]
Religiosity	.00	.05	-.08	.938	[-.11, .10]
Replication Dataset 2					
Binding Values	.24	.06	4.16	< .0001	[.13, .35]
Political Orientation	-.02	.03	-.60	.55	[-.08, .04]
Gender	-.19	.09	-2.09	.037	[-.36, -.01]
Religiosity	-.04	.02	-1.54	.12	[-.08, .01]

Note. Study 1 (*N* = 459), Replication Dataset 1 (*N* = 249), Replication Dataset 2 (*N* = 788). All 95% CIs are for the beta-estimates.

D: Individualizing values and Implicit Causality Object-Bias

We investigated whether individualizing values also predict a subject- or object-bias in the implicit causality task. Previously, individualizing values were found to be positively associated with perpetrator blame (Niemi & Young, 2016). This association was notably weaker than the associations between binding values and judgments of victims as blameworthy and responsible. Therefore, we did not have strong expectations regarding the implicit causality behavior of participants high in individualizing values. Nevertheless, increased selection of the subject for harm/force verbs would be consistent with the prior findings of increased perpetrator blame. To investigate this, 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 individualizing 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.

In Study 1, there was no significant main effect of individualizing values (*b* = -.02, *SE* = .08, *Z* = -.27, *p* > .05, 95% CI = [-.18, .13]), no significant main effect of verb type (*b* = .88, *SE* =

.52, *Z* = 1.72, *p* > .05, 95% CI = [-.13, 1.90]), and no significant interaction (*b* = .06, *SE* = .07, *Z* = .85, *p* > .05, 95% CI = [-.08, .20]). There was a small but significant interaction in Replication Dataset 1 (*b* = .23, *SE* = .09, *Z* = 2.44, *p* = .01, 95% CI = [.04, .41]), but there were no significant main effects (*p*s > .05). Despite this significant interaction effect, there was still no effect of individualizing values on the propensity to select the object *relative to* the subject as the referent for the subset of harm/force verbs (*p* > .05) or the subset of neutral filler verbs (*p* > .05), when these verb types were modeled separately. In Replication Dataset 2, there was no significant main effect of individualizing values (*b* = -.08, *SE* = .06, *Z* = -1.49, *p* > .05, 95% CI = [-.19, .03]) and no significant interaction between individualizing values and verb type (*b* = .07, *SE* = .04, *Z* = 1.81, *p* > .05, 95% CI = [-.01, .14). There was, however, a small but significant main effect of verb type (*b* = .72, *SE* = .35, *Z* = 2.06, *p* = .04, 95% CI = [.04, 1.40]). Therefore, across the three datasets, it is reasonable to conclude that binding values, not individualizing values, are associated with the object-bias (for harm/force verbs and not neutral filler verbs).