Implicit measurement of motivated causal attribution

Laura Niemi (lauraniemi@fas.harvard.edu)[1], **Joshua Hartshorne (jharts@mit.edu)**[2,3]
Tobias Gerstenberg (tger@mit.edu)[2] & **Liane Young (liane.young@bc.edu)**[3]
[1]Department of Psychology, Harvard University, Cambridge, MA 02138
[2]Massachusetts Institute of Technology, Cambridge, MA 02139
[3]Department of Psychology, Boston College, Chestnut Hill, MA 02467

Abstract

Moral judgment often involves pinning causation for harm to a particular person. Since it reveals "who one sides with", expression of moral judgment can be a costly social act that people may be motivated to conceal. Here, we demonstrate that a simple, well-studied psycholinguistic task (implicit causality) can be leveraged as a novel implicit measure of morally relevant causal attributions. Participants decided whether to continue sentences like *"Amy killed Bob because..."* with either the pronoun *he* or *she*. We found that (1) implicit causality selections predicted explicit causal judgments, (2) selecting the object (victim) for harm/force events (e.g., *kill, rape*) predicted endorsement of moral values previously linked to victim-blame, and (3) higher hostile sexism predicted selecting the female as the cause in male-on-female harm/force. The implicit causality task is a new measure of morally motivated causal attribution that may circumvent social desirability concerns.

Keywords: implicit cognition; causation; psycholinguistics; moral psychology; implicit causality; semantics

Introduction

Blame and condemnation is often placed on the perceived cause of a negative outcome (Malle, Guglielmo, & Munroe, 2014; Cushman, 2008). However, while we often talk loosely about *the* cause of an event (Hilton, 1990), the truth is far more complex. For instance, we might say that Brutus slew Caesar because he despised him. However, Brutus might explain Caesar's death by referring to a different, prior cause: "Caesar was ambitious, so I slew him." In fact, there are many points along the causal chain that links the Big Bang to Caesar's death, any of which could conceivably be referenced in an explanation.

Nonetheless, people tend to focus on certain causes at the expense of others. For example, people typically focus on causes that are more proximal to the event, ignoring those that are more distal (cf. White, 1992; Hilton, McClure, & Sutton, 2010). Both Caesar's birth and his presence in the Forum on the Ides of March were causally necessary for the murder, but there is something unsatisfactory about the explanation that *Brutus slew Caesar because Caesar was born*. Furthermore, when perceiving causal events, humans tend to focus on powers and ignore liabilities (White, 2006, 2007). For example, classic Michottean perception of causation involves asymmetric attention to power over liability: even though the "launched" ball could be construed as *stopping* the other ball, people reliably perceive a "launching" rather than a "stopping" event (Michotte, 1963; Mayrhofer & Waldmann, 2014; White, 2006).

Since causal powers are mapped to "agents" and causal liabilities to "patients" (White, 2006), this suggests that people will construe causality asymmetrically across the agent-patient dyad during social and moral reasoning. In one sense, this seems obvious: powers of the "agent" (e.g., the ability to kill) seem much more likely to be deemed causal than liabilities of the "patient" (i.e., the ability to be killed). Brutus could not have killed an immortal, but Caesar's mortality is unlikely to be considered a candidate explanation (*Brutus killed Caesar because Caesar is mortal*). However, agents are more than their physical powers and liabilities; actions are imbued with social and moral meaning that cannot be boiled down to physical descriptions. While humans may prefer certain kinds of explanations and focus on certain kinds of causes, how the "causal asymmetry" plays out in social-moral domain is far from clear (White, 2006).

People can and do discuss distal causes, patients, and liabilities during evaluation of morally relevant events (Heider & Simmel, 1944). Thus, understanding human blame assignment and condemnation behavior requires rich theories of causation (e.g. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Wolff, 2007). Deepening our understanding of the representation of causation in moral judgment, in turn, is crucial for both our theoretical understanding of moral cognition as well as for public policy and organizational planning.

However, the social importance of moral reasoning makes it difficult to study: subjects may be motivated to give socially desirable answers or conceal their thoughts and attitudes (Alicke, 2000; Banaji & Heiphetz, 2010; Fazio & Olsen, 2003). Prior work shows that attributions often disfavor people representing social groups that participants view in an adversarial manner and favor people representing social groups with whom they are allied (Morgan, Mullen, & Skitka, 2010). Thus, causal attributions can signal people's *specific* personal alliances and hostilities. In public discourse, collaborative settings such as the workplace, as well as adversarial situations, revealing partiality can be socially detrimental.

A key challenge is to find tasks that implicitly measure morally relevant judgments in such a way that circumvents guarding or concealment. There are now multiple well-validated implicit measures that track people's attitudes about and associations between concepts – most notably the Implicit Association Test (IAT) (Nosek, Banaji, & Greenwald, 2002; Banaji & Heiphetz, 2010; Fazio & Olsen, 2003; Nock et al., 2010; Nosek et al., 2002). However, so far, there are no good implicit measures of moral reasoning and blame assignment *per se*. In this study, we show that a well-understood phenomenon from psycholinguistics – implicit causality – presents just such a measure.

Implicit Causality

People have reliable expectations about why some events happened. Most people expect the sentence

1. John frightened Amy because...

to continue with a reference to the sentence subject, John (e.g., *he was carrying a gun*), but

2. John feared Amy because...

to continue with a reference to the sentence object, Amy (e.g., *she had a bad temper*).[1]

Verbs' tendencies to lead people to select the subject or the object are called their "implicit causality biases" on the basis that they reflect automatic intuitions about the (most important) cause of the event in question (Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Rudolph & Forsterling, 1997; VanBerkum, Koornneef, Ottena, & Nieuwland, 2007).There are many subject-biased verbs like *frightened* and object-biased verbs like *feared*, as well as verbs that produce no reliable trend in either direction (Pickering & Majid, 2007; Rudolph & Forsterling, 1997; Hartshorne & Snedeker, 2012; Ferstl, Garnham, & Manouilidou, 2011). The implicit causality task used here involves simply asking participants to decide whether sentences like (1) or (2) continue with the pronoun *he* or *she* (Hartshorne, 2013).

Interest in determining how exactly implicit causality selections map onto causal cognition has produced a large and contentious literature (Bott & Solstad, 2014; Hartshorne & Snedeker, 2012; Hartshorne, 2013; Pickering & Majid, 2007; Brown & Fish, 1983; Rudolph & Forsterling, 1997). This work – which we return to in the Discussion – has largely focused on determining what the differences are between verbs that result in different implicit causality selections (e.g., *frighten* vs. *fear*). Here, we show that *individual variation* in selections provides a window into individuals' (moral) beliefs, values, and reasoning processes.

Explicit and Implicit Judgments

We compare implicit causality bias and explicit judgments about the causal responsibility of the participants in the event, particularly, beliefs indicative of a view that victims of violence "had it coming" and that such events are "victim-precipitated". While a view that a victim "probably had it coming" is likely to be perceived as relatively "safe" to endorse explicitly in an anonymous online survey – which is what we used – these are precisely the sorts of beliefs that people may be motivated to conceal in a non-anonymous context (e.g., the workplace or other setting in which impartiality is the social norm; also, adversarial situations).

Here, we compared implicit causality bias to judgments that the agent (sentential subject) was necessary and sufficient for what happened, and the patient (sentential object) allowed, controlled and deserved the outcome. We expected a preference to select the object ("object-bias", henceforth) across events of harm and force to predict explicit beliefs that agents were less necessary and sufficient, and that patients were more likely to have allowed, controlled and deserved the events. If implicit causality selections predict explicit ratings of agents' and patients' causal capacities – judgments closely tied to moral judgment (Alicke, 1992; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015) – then this provides initial support for the instrument's utility as an implicit measure of morally relevant causal attributions.

Beliefs and Values

We expected that participants would be more likely to demonstrate an object-bias for harm and force verbs when shifting moral responsibility off the person in the subject position and onto the person in the object position would align with their personal beliefs and values. In the current research, we test two case studies of such motivated causal attribution related to (a) moral values, and (b) sexism.

Our predictions related to moral values align with prior work showing that *binding values* – a cluster of highly intercorrelated moral values that censure disloyalty, disobedience to authority, and sexual/spiritual "impurity" (Graham et al., 2011) robustly predict attributions of responsibility and blame to victims (Niemi & Young, in press). If implicit causality biases to objects of events of harm and force are more likely in people higher in binding values, this represents evidence of convergent validity in support for our claim that the implicit causality task taps morally motivated causal attributions.

We also examine the capacity of the implicit causality task to serve as an implicit measure of people's hostility toward a particular social category, using the test case of hostility toward women. To this end, we manipulated the gender of the subject and object in the implicit causality prompts (e.g., "John" *verbed* "Mary"; and vice versa), and measured participants' hostile sexism (Glick & Fiske, 1996). We expected participants who were higher in hostile sexism, which involves antagonistic attitudes toward women, to be more likely to select the object as causal in the implicit causality task only when men were presented as harming women, and not when women were presented as harming men. That is, implicit causality selections should reflect a view of violence against women as more "victim-precipitated" in people high in hostile sexism. Notably, on our account, sexism should only be correlated with object-bias for verbs of harm and force when men are in the subject position and women are in the object position, and not vice versa; whereas binding values should be correlated with object-bias for verbs of harm and force regardless of gender of subject and object.

Method

459 participants were recruited via Amazon Mechanical Turk ($M_{age} = 37.25, SD_{age} = 31.39$; 207 female, 247 male, 5 selected other or missing). 314 additional individuals failed

[1] "Implicit causality" has been used to refer to two different phenomena (Hartshorne, 2013). One involves a tendency for verbs to lead people to refer to the sentence subject or object following the *because* conjunction; this is the phenomenon investigated here. The other involves a tendency for verbs to trigger intuitions about covariation; see Discussion.

attention checks ($n = 189$); didn't complete the study or previously took a related study ($n = 125$), and were excluded from analyses.

Implicit Causality Task

Participants viewed 24 randomized prompts in the format "*[Agent] [verb]*ed *[Patient]* because" and were asked to "Please select which word you think would follow." They were offered the choices "he" or"she" (counterbalanced pronoun order across items).

We varied agent and patient gender between participants, presenting half of the sample with male agents and female patients, and vice versa for the other half. Verbs included 12 conveying harm and force ("harm/force verbs" henceforth); and 12 filler verbs with ranging causal biases (Bott & Solstad, 2014; Hartshorne & Snedeker, 2012). Verbs and the probability of selecting the object ("object-bias") are shown in Table 1.

Measures of Beliefs about Agents and Patients

After completing the implicit causality task, participants were instructed to "Consider a hypothetical event:" and were again presented with the 24 events they had seen during the implicit causality task but this time without the "because" connective (e.g., "George impressed Julie."). Each event was followed by instructions to "Weigh the following possibilities:" and a series of items in the following order:

1. Agent Unnecessary: "Would *[patient]* have been *[verbed]* by someone else?"
2. Agent Sufficient: "Would *[agent] [verb]* someone else?"
3. Patient Control: "Did *[patient]* have control over the occurrence of the event?"
4. Patient Allowing: "Did *[patient]* let the event happen?"
5. Patient Desert: "Could *[patient]* have deserved the event?"

Participants responded using sliding scales anchored at 0 = "Definitely No"; 50 = "Unsure"; 100 = "Definitely Yes". We created variables capturing "Agent Contribution" by averaging the Agent Unnecessary ratings (reverse-coded) and Agent Sufficient ratings (*Cronbach's alpha*=.78); and "Patient Contribution" by averaging the Patient Control, Patient Allowing, and Patient Desert ratings (*Cronbach's alpha*=.79).

Moral Values Questionnaire

Moral values in the five foundations (caring, fairness, ingroup loyalty, authority, and purity) were assessed using the 30-item Moral Foundations Questionnaire (MFQ) (Graham et al., 2011)."Individualizing values" represent the extent of endorsement of caring and fairness values. "Binding values" represent the extent of endorsement of ingroup loyalty, authority and purity values. Participants also provided demographic information including politics, gender and religiosity.

Ambivalent Sexism Inventory

Sexism was assessed with the Ambivalent Sexism Inventory (ASI) (Glick & Fiske, 1996). The ASI measures extent of agreement with 22 statements about women and men in society and allows for the calculation of a hostile sexism score and a benevolent sexism score. Hostile sexism captures antagonistic attitudes toward women and toward women's pursuit of equality, whereas benevolent sexism captures traditional stereotypes about women as requiring protection by men. In order to directly index adversarial attitudes toward women, we used the hostile sexism subscore.

Results

Implicit Causality and Beliefs about Agents and Patients We hypothesized that the implicit causality task provides a window into beliefs that victims of violence "had it coming" – which people may be motivated to conceal in many contexts. Regression analysis with Agent and Patient Contribution entered simultaneously confirmed that object-bias in the implicit causality task was negatively predicted by Agent Contribution ratings ($\beta = -.176, p < .001$) and positively predicted by Patient Contribution ratings ($\beta = .240, p < .001$) for harm/force verbs (Model R^2 =.13).[2] The fact that perceptions of patients as having the capacity to control, allow and deserve events of harm and force traded off with perception of agents as necessary and sufficient in the model ($r = -.483, p < .001$) indicates that the nature of the implicit causality task, which forces a choice between the agent or patient, aligns well with causal intuitions. Patient Contribution ratings were relatively low for events of harm and force[3], which aligns with intuitive jurisprudence: perpetrators of violence and imposition are typically viewed as more causally responsible than victims.

In sum, increased beliefs that patients controlled, allowed and deserved events and decreased beliefs that agents were necessary and sufficient were associated with increased likelihood of selecting people in the object position (patients) rather than people in the subject position (agents) as causal in the implicit causality task for events involving harm and force (see Fig.1a-b). As an illustrative example, ratings of the likelihood that patients *deserved being killed* were higher for participants who continued the sentence "Bob killed Amy because" with the pronoun referring to the patient – *she* ($M = 26.10, SEM = 1.89$) rather than the agent – *he* ($M = 19.62, SEM = 1.79$).

Implicit Causality and Moral Values Prior work has found that greater endorsement of binding values – loyalty, respect for authority and preservation of purity – predicts greater attribution of blame to victims (Niemi & Young, in press). Accordingly, we expected greater endorsement of binding values to correlate with object-bias for harm/force verbs in the implicit causality task. As hypothesized, binding values were correlated with object-bias for harm/force verbs ($r = .294, p < .001$), but not filler verbs ($r = .07, p = .14$). To illustrate, binding values positively predicted the likelihood

[2]Object-bias for filler verbs was positively predicted by Patient Contribution ratings ($\beta = .169, p < .001$), but not Agent Contribution ratings ($p = .54$).

[3]Patient Contribution ratings for harm/force verbs ($M = 31.5, SEM = .77$) were significantly lower than for filler verbs ($M = 55.42, SEM = .62, t(458) = 32.27, p < .001$). See Table 1 for means.

Table 1: Average implicit causality object-bias (OB), Agent Contribution, and Patient Contribution ratings.

Verb	OB	Agent	Patient
Harm/Force			
Clobbered	0.55	62.76	29.26
Coerced	0.30	56.96	41.41
Enslaved	0.36	66.39	22.88
Forced	0.39	61.11	29.65
Influenced	0.31	51.78	58.70
Killed	0.53	67.08	19.66
Manipulated	0.29	57.52	40.69
Raped	0.30	70.15	15.86
Robbed	0.26	66.05	19.92
Stabbed	0.50	66.00	22.47
Strangled	0.54	66.09	22.70
Tempted	0.31	52.75	54.80
Fillers			
Approached	0.44	50.63	50.74
Confused	0.29	51.36	40.53
Congratulated	0.88	46.75	65.67
Delighted	0.30	52.49	62.26
Impressed	0.20	52.90	58.71
Observed	0.62	51.52	41.38
Praised	0.85	49.04	63.94
Punished	0.74	57.34	43.96
Quoted	0.61	51.60	47.38
Skipped	0.60	54.34	37.54
Thanked	0.83	51.45	65.25
Transported	0.73	51.31	67.67

that participants would select the referent to the patient rather than the agent following "*[Agent]* killed *[Patient]* because" (*Odds Ratio (OR)*=1.41, p<.001).

Also as hypothesized, correlations between binding values and object-bias for harm/force verbs were maintained across gender frames, no matter whether male-verbed-female ($r = .399, p < .001$), or female-verbed-male ($r = .201, p = .002$).[4] Individualizing values were uncorrelated with object-bias for harm/force verbs and filler verbs (p's>.46).

To determine the contribution of binding values above and beyond associated beliefs about agents and patients, we entered these variables simultaneously into regression analyses.[5] Analyses revealed that binding values ($\beta = .206, p < .001$), Patient Contribution ($\beta = .195, p < .001$), and Agent Contribution ($\beta = -.151, p = .002$) significantly predicted object-bias for the harm/force verbs (Model R^2=.17).

In sum, increased binding values – in addition to beliefs that patients allowed, controlled and deserved the harm/force events, and that agents were not necessary or sufficient – predicted object-bias for harm/force verbs (see Fig.1c).

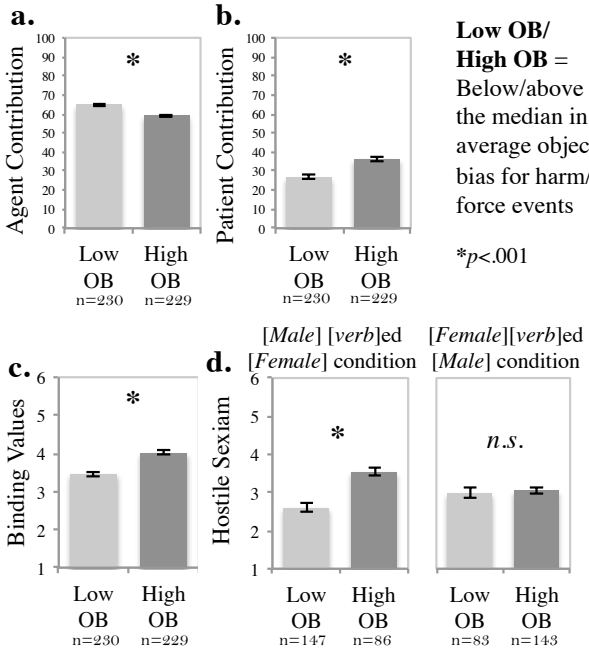

Figure 1: Primary results: Participants who were more likely to select the object as causal in the implicit causality task for harm/force events (e.g., selected "she" not "he" for events such as "Bob killed Amy because ... he or she?") showed (a) reduced ratings of agents' causal contributions, (b) increased ratings of patients' causal contributions, and (c) higher binding values. Participants who were more likely to select the object as causal for harm/force events when agents were male and patients were female showed higher hostile sexism (d, left panel), but not when agents were female and patients male (d, right panel).

Implicit Causality and Sexism Supporting hypotheses, hostile sexism was correlated with object-bias for harm/force verbs when male-verbed-female ($r = .439, p < .001$; Fig.1d, left panel), but *not* when female-verbed-male ($r = .05, p = .45$; Fig.1d, right panel). To illustrate, higher hostile sexism indicated a higher likelihood of selecting "she" following "Bob killed Amy because ... he or she?" ($OR = 1.94, p < .001$), but had no predictive value for selections following "Amy killed Bob because ... he or she?" ($OR = .927, p < .54$). These results indicate that a view of men's violence toward women as "victim-precipitated" in people high in hostility toward women is measurable with the implicit causality task.

Replications

The above study was replicated twice as part of two follow-up studies. We again found significant correlations between binding values and implicit causality object-bias for harm/force events [Replication 1: n=788 ($r = .137, p < .001$); Replication 2: n=249 ($r = .287, p < .001$)] and significant correlations between hostile sexism and implicit causality object-bias for harm/force events when men were sentence subjects and women sentence objects [Replication 1, n=410 ($r = .375, p < .001$); Replication 2, n=121 ($r = .459, p < .001$)], but *not* when women were sentence subjects and men sentence objects ($p's > .84$).

[4]Correlations were significantly different ($Z = 2.32, p = .02$).

[5]Notably, and aligning with prior work showing that binding values predict increased attribution of responsibility and blame to victims and reduced counterfactual focus on perpetrators (Niemi & Young, in press), binding values were positively correlated with Patient Contribution ($r = .276, p < .001$) and negatively correlated with Agent Contribution ($r = -.226, p < .001$) for harm/force verbs.

Discussion

The implicit causality task is a novel measure of motivated causal attribution. Average object-bias in the task (i.e., likelihood of selecting a pronoun referring to the person in the object position to follow *"X verbed Y because"*) positively correlated with beliefs that patients allowed, controlled, and deserved harm/force events. Implicit causality object-bias for harm/force events was also positively correlated with endorsement of binding values, which were previously found to predict victim blaming (Niemi & Young, in press). Finally, implicit causality object-bias for harm/force events when men were sentence subjects and women were sentence objects was positively correlated with hostile sexism. This demonstrates that more specific patterns of hostility toward people in particular social categories (here, women) are traceable with the implicit causality task. We replicated these results twice.

The relationship between implicit causality selections and ratings of agents' and patients' contributions to harm and force indicates that the implicit causality task has predictive validity as an index of people's causal attributions across the agent-patient dyad. This finding has implications for an influential theory of moral cognition – "dyadic morality" – which argues that the *agent-harms-patient* template represents the core structure of immoral events (Gray, Young, & Waytz, 2012; Schein & Gray, 2015; Gray, Schein, & Ward, 2014).

On this account, a universal feature of moral judgment is a tendency for people to justify their moral judgments by identifying harmed victims and asymmetrically loading causality and blame onto harmful perpetrators (Gray et al., 2014; Schein & Gray, 2015). The current results indicate that there is systematic diversity in how people attribute causation across the dyad for events that involve harm and force.

Prior work has argued that implicit causality biases do not reliably predict explicit causal judgments, at the verb-specific level of analysis (Hartshorne, 2013), based on examining the relationship between implicit causality selections and people's ratings of covariation (i.e., ratings of subject as "the kind of person who verbs people" and the object as the "kind of person people verb"; Brown & Fish, 1983). In line with this prior work, we found that participants' ratings of agents' sufficiency and necessity, which involved a counterfactual framing of similar items in our study, did not pattern with implicit causality ratings for *individual verbs*. The capacity of the implicit causality task to tap into people's causal reasoning about agents and patients was revealed when averaging *across* verbs of harm and force. This indicates that the task is particularly well-suited to reveal one's general construal of the nature of dyadic human action involving harm and imposition, namely, how much one views harm as sourced in autonomous agents versus compelled by precipitating patients.

The positive relationship between object-bias in the implicit causality task and binding values aligns with prior findings that binding values predict victim blame via increased attributions of responsibility to victims (Niemi & Young, in press). Why might binding values predict attributions to patients at the implicit and explicit level? Unlike *individualizing values* which promote unconditional care and straightforwardly define agents' harm to patients as *always wrong*, binding values – loyalty, respect for authority, concern for purity – involve moralization of norms which sometimes *require* harm (Graham et al., 2011; Niemi & Young, in press). For example, loyalty to the ingroup may entail harming people belonging to the outgroup; harming or banishing a group member who has brought dishonor on the group may be compatible with moral valuation of purity (e.g., honor killing). Thus, people who highly endorse binding values may deviate from the typical pattern of moral judgment focused on events fitting an *agent-harms-patient* template (Gray et al., 2012; Schein & Gray, 2015). Indeed, aligning with the prior work focused on moral judgment (Niemi & Young, in press), here we found that people higher in binding values were those more likely to show an "inverted" pattern of attribution for events of harm and force whereby patients were selected as causal over agents.

We also found evidence that the implicit causality task tracks more specific motivated attributions. Hostile sexism was correlated with object-bias for events of harm and force when women were in the object position (and men in the subject position), but not when men were in the object position (and women in the subject position). This result indicates that hostility toward a specific social category (here, women) meaningfully predicts implicit attributions of causality across events of harm and force to representatives of that social category in the position of "harmed-patient" – in addition to broader moral commitments (i.e., binding values). Future work will investigate the influence of manipulations of the implied race of the sentence subject and object together with individual differences in racism on implicit causality for events of harm and force. Other work will explore the extent to which viewing agent-patient events as induced by patients serves post hoc justification of moral condemnation or exculpation of agents with whom one feels at risk of being identified (Heider, 1958; Morgan et al., 2010) by experimentally manipulating group membership (e.g., Masten et al., 2010).

Prior work has demonstrated that implicit measures can tap into people's hidden biases (Fazio & Olsen, 2003). The IAT, for example, is a powerful implicit instrument because even if participants infer the experimenter's aim to measure negative attitudes toward a target and attempt to alter their task behavior, they are not likely to be successful. Nevertheless, researchers may be interested in measuring attitudes toward a target more covertly. One advantage of the implicit causality task is that it can be embedded with distractor items, potentially disguising experimenters' aims. And since individual items are answered very quickly, demand characteristics and participant bias can be kept low in this manner without making the task excessively long. Moreover, the ease of administrating the implicit causality task will make it simple to combine with functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and other measures of neural activity in future investigations of the neurobiological mecha-

nisms involved in social judgment and causal attribution. The implicit causality task can also be completed with a pencil and paper, or administered verbally. Thus, it has the capacity to tap attitudes via causal attributions in a broad range of populations, including people unfamiliar with or physically incapable of using technological devices, and people who are not literate, including young children. Cross-linguistic investigations with the implicit causality task are also suitable (cf. Hartshorne, Sudo, & Uruwashi, 2013), which will enable researchers to shed light on cultural differences in motivated causal attribution. Ongoing work is now focused on characterizing the extent of the *implicitness* of the implicit causality task, involving, in part, manipulating task instructions and probing participants' thoughts on the purpose of the task.

Conclusion

We have shown that responses to the implicit causality task predict people's general beliefs about agents' and patients' causal contributions to events involving violence and imposition, as well as their higher-level moral values and social attitudes. The implicit causality task provides a versatile tool for future research into motivated causal attribution.

Acknowledgments The authors gratefully acknowledge the helpful comments of Steven Pinker, Fiery Cushman, Jorie Koster-Hale, Jonathan Phillips, Alek Chakroff, the ECOM Research Group, the Events in Language Reading Group at Harvard, and the Morality Lab at Boston College.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556.

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*(6), 790-812.

Banaji, M. R., & Heiphetz, L. (2010). Handbook of social psychology. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), (p. 348-388). New York: Wiley.

Bott, O., & Solstad, T. (2014). Psycholinguistic approaches to meaning and understanding across languages (studies in theoretical psycholinguistics). In B. Hemforth, B. Mertins, & C. Fabricius-Hansen (Eds.), (p. 213-251). Cham: Springer.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *17*, 237-273.

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

Fazio, R. H., & Olsen, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in english: A corpus of 300 verbs. *Behavior Research Methods*, *43*(1), 124-135.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459-464.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Glick, P., & Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491-512.

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., Schein, C., & Ward, A. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600-1615.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124.

Hartshorne, J. K. (2013). What is implicit causality? *Language, Cognition and Neuroscience*, *29*(7), 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*(10), 1474–1508.

Hartshorne, J. K., Sudo, Y., & Uruwashi, M. (2013, feb). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, *60*(3), 179–196. doi: 10.1027/1618-3169/a000187

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*(3), 383–400.

Malle, B. F., Guglielmo, S., & Munroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Mayrhofer, R., & Waldmann, M. R. (2014, Sep). Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, *132*(3), 485-490. doi: 10.1016/j.cognition.2014.05.013

Michotte, A. E. (1963). *The perception of causality*. New York: Basic Books.

Morgan, G. S., Mullen, E., & Skitka, L. J. (2010). When values and attributions collide: Liberals and conservatives' values motivate attributions for alleged misdeeds. *Personality and Social Psychology Review*, *36*(9), 1241-1254.

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

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511-517.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research and Practice*, *6*(1), 101-115.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*(5), 780-788.

Rudolph, U., & Forsterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin*, *121*(2), 192-218.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147-1163.

VanBerkum, J. J. A., Koornneef, A. W., Ottena, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158-171.

White, P. A. (1992). Causal powers, causal questions, and the place of regularity information in causal attribution. *British Journal of Psychology*, *83*(2), 161–188.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147.

White, P. A. (2007). Impressions of force in visual perception of collision events: A test of the causal asymmetry hypothesis. *Psychonomic Bulletin & Review*, *14*(4), 647-652.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.