Journal of Experimental Social Psychology 90 (2020) 103990

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Implicit evaluations of moral agents reflect intent and outcome☆,☆☆

Benedek Kurdi*, Amy R. Krosch, Melissa J. Ferguson

Department of Psychology, Cornell University, Ithaca, NY, United States of America

ARTICLE INFO

Keywords:
Implicit Association Test
Implicit evaluations
Implicit social cognition
Morality
Propositional theories
Theory of mind

ABSTRACT

Explicit (directly measured) evaluations of moral agents reflect both the externally observable consequences of actions and inferences about the agent's hidden mental states: Negative outcomes without negative intent (e.g., someone getting killed accidentally) and negative intent without a negative outcome (e.g., a failed attempt to kill someone) are each sufficient for negative explicit evaluations of a moral agent to emerge. Across two studies (final N = 826; Study 2 preregistered), we newly investigated implicit (indirectly measured) evaluations of moral agents, as assessed by an Implicit Association Test (IAT). Study 1 included 3 between-participant conditions: accident (negative outcome + positive intent), attempt (positive outcome + negative intent), and harm (negative outcome + negative intent), each compared to a harmless (positive outcome + positive intent) control. Study 2 had a 2-by-2 design, in which outcome (positive vs. negative) and intent (positive vs. negative) were manipulated orthogonally, with targets in each condition compared to a neutral control whose actions did not carry moral implications. Mirroring prior findings obtained using explicit measures, implicit evaluations of moral agents tracked both manifest outcomes (e.g., someone falling from a bridge) and inferences about latent mental states (e.g., the intent to let someone fall off a bridge) in both paradigms. These results are difficult to reconcile with dual-process theories positing that implicit evaluations arise from low-level associative learning but they are readily accounted for by propositional theories according to which implicit evaluations are sensitive to high-level inferential reasoning.

1. Introduction

In determining whether a person has committed a crime, and if so what punishment they deserve for it, most legal systems the world over rely on a combination of two factors: the outcome of the person's actions, directly observable in the external world, as well as their hidden mental states (such as beliefs, thoughts, and desires) when performing those actions. Specifically, a person causally responsible for harming someone else, such as fatally injuring a pedestrian in a car accident, can be convicted of a crime even in the complete absence of any intent to harm. Conversely, the intent to harm can also, in and of itself, be sufficient to warrant punishment. For instance, attempted murder using ineffective means, such as unknowingly hiring an undercover police officer as a hit man, constitutes a crime in the absence of any harmful consequences in the world. Of course, in most cases of criminal wrongdoing, adverse real-world consequences and malicious intent are both present to varying degrees. In such cases, the same action can be subject to more severe sanctions if it was intended rather than unintended (e.g., murder vs. manslaughter), and the same intent can be subject to more severe sanctions if the criminal act was successful rather than unsuccessful (e.g., achieved vs. attempted murder).

When laypeople are asked to judge the culpability or character of moral actors using carefully crafted scenarios in the lab, their responses seem to reflect principles that are surprisingly similar to the principles applied in a more codified manner in the institutionalized setting of legal proceedings. Specifically, in studies on moral judgment, participants appear to rely on both outcomes (real-world consequences) and intent (unobservable mental states) in assigning blame (or praise) for morally relevant actions and in evaluating moral actors (Cushman, 2008; Cushman, Sheketoff, Wharton, & Carey, 2013; Knobe, 2005; Young, Campodron, Hauser, Pascual-Leone, & Saxe, 2010; Young, Cushman, Hauser, & Saxe, 2007). In an influential set of experiments, Young et al. (2007) found negative outcomes and malicious intent to be independently sufficient for negative evaluations of a moral actor to emerge. For instance, relative to a control person who had performed a morally innocuous act, participants expressed reduced positivity

☆ This paper has been recommended for acceptance by Rachel Barkan.

☆☆ Parts of this research were presented at the Social Cognition Preconference of the 21st Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA, in February 2020. We thank Fiery Cushman for helpful discussions about the study design.

* Corresponding author at: Department of Psychology, Cornell University, Ithaca, NY 14850, United States of America.
E-mail address: bk493@cornell.edu (B. Kurdi).

https://doi.org/10.1016/j.jesp.2020.103990
Received 11 December 2019; Received in revised form 19 March 2020; Accepted 26 March 2020
0022-1031/ © 2020 Elsevier Inc. All rights reserved.

toward a person who was described as having accidentally let a person die because he erroneously believed that an unsteady bridge was steady. Participants also showed marked negativity toward a person described as unsuccessfully attempting to kill someone by letting them cross a steady bridge that they erroneously believed was about to collapse. Evaluations were most markedly negative in a third scenario where negative outcome and malicious intent were both simultaneously present, i.e., the murder was successful.

In moral psychology work, including all the studies mentioned above, judgments about moral actors and their actions are routinely elicited using self-report measures, which provide ample time for intentional processes of elaboration to unfold. However, as demonstrated by a large and rapidly growing body of research since the 1980s, social evaluation also readily occurs automatically, that is, with relatively low levels of intention, awareness, or control (Bargh, 1994; Devine, 1989; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995). In fact, dual-process theories, which have dominated social cognition research over the past decades, make the central assumption that explicit (directly measured) and implicit (indirectly measured) cognition arise from fundamentally different computations and predict different kinds of behavior (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Indeed, multiple meta-analyses have provided evidence for the unique predictive power of implicit cognition, above and beyond explicit cognition, including in consequential real-world settings (e.g., Cameron, Brown-Iannuzzi, & Payne, 2012; Kurdi et al., 2019). As such, the question of whether implicit moral evaluation, like explicit moral evaluation, is sensitive to both outcome and intent is not only of basic theoretical interest but may also have implications for explaining real-world behavior.

Previous research on implicit social cognition has produced considerable evidence for the responsiveness of indirectly measured evaluations to behaviors with moral implications (e.g., Cone & Ferguson, 2015; Mann & Ferguson, 2015; Moran & Bar-Anan, 2013; Peters & Gawronski, 2011; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2006). For instance, Rydell et al. (2006) have demonstrated that implicit evaluations of a novel target respond to brief descriptions of morally virtuous (e.g., “Bob fought against a discriminatory law that made renting difficult for minorities”) and morally repugnant (e.g., “Bob continually yells at his wife in public”) behaviors in the theoretically expected direction. Peters and Gawronski (2011) have provided evidence for the sensitivity of implicit evaluations to similar behavioral descriptions (e.g., “Mike lent money to a friend in financial trouble” vs. “Mike cheated during a poker game”). Work by Ferguson and colleagues has revealed flexibility in the updating of implicit evaluations in response to highly diagnostic, i.e., morally relevant, information. For instance, Cone and Ferguson (2015) have shown that a single piece of diagnostic information, such as someone having mutilated a small defenseless animal, can reverse implicit evaluations created from 100 behavioral statements of the kind used by Rydell et al. (2006).

This body of work offers ample evidence that morally relevant behaviors can play a major role in the emergence and updating of implicit evaluations. However, the studies in question were designed to inform theoretical issues such as differences between explicit and implicit evaluations in sensitivity to counterattitudinal information and in doing so they confounded the two variables of central interest here: intent and outcome. For instance, in the studies by Cone and Ferguson (2015) it is unclear whether the target was evaluated negatively on implicit measures as a result of a highly negative outcome (e.g., being causally responsible for harming a defenseless animal), malicious intent (e.g., the mere desire to harm a defenseless animal), or a combination of both.

A separate set of experiments have demonstrated that a causal connection between a target and a valenced event can lead to stronger implicit evaluations than mere association between the two (Cone & Ferguson, 2015; Hughes, Ye, Van Dessel, & De Houwer, 2019; Kurdi, Morris, & Cushman, 2020; Moran, Bar-Anan, & Nosek, 2015, 2016; but see Moran & Bar-Anan, 2013). For instance, Hughes et al. (2019) exposed participants in all experimental conditions to the same pairings of a human target with valenced words (such as “fantastic” vs. “horrible”) but used verbal instructions to manipulate the perceived relationship between the two. Specifically, in one condition the human target was described as having caused the appearance of valenced words, in a second condition as predicting the appearance of valenced words, and in a third condition as being unrelated to the appearance of valenced words. Implicit evaluations were most strongly in line with the valence suggested by the pairings in the cause condition, followed by the prediction condition, and finally by the unrelated condition. Similarly, in an experiment by Cone and Ferguson (2015), a target person was described either as having committed a crime (causal condition) or as attending a high school where someone else had committed a crime (association condition). Implicit evaluations of the target were found to be negative only in the causal but not in the association condition.

Jointly, these studies suggest that implicit social cognition can be sensitive to the relationship between causes and effects, with a causal connection increasing the strength of implicit evaluations relative to mere association (see also Moran et al., 2015). However, it is unclear whether in these studies, and in the related studies by Moran and colleagues, participants made spontaneous inferences about the human targets' hidden mental states in addition to representing causal relationships. For instance, in the causal condition of the study by Hughes et al. (2019), participants could have concluded that the target not only caused valenced events in the physical sense of the word but may also have intended for this kind of outcome to occur. Kurdi et al. (2020) have started disambiguating these findings by showing that physical objects (e.g., colored shapes) causally responsible for a valenced outcome (e.g., a machine dispensing a diamond) are subject to more extreme implicit evaluations than objects merely co-occurring with the same valenced outcome. This result suggests that inferences about an actor's hidden mental states are not necessary for an influence of causal relationships on implicit evaluations to emerge. At the same time, the experiments by Kurdi et al. (2020) included no human actors and, as such, leave open the question of whether similar effects would also be obtained in the context of human targets, which are by definition characterized by their ability to experience mental states (Gray, Gray, & Wegner, 2007).

To summarize, existing social cognition work has demonstrated the sensitivity of implicit evaluations to morally relevant behaviors and causal relationships. However, studies involving morally relevant behaviors have confounded the effects of intent and outcome—the two major building blocks of moral judgment—on implicit evaluations. A second set of studies have established that being causally responsible for (rather than merely associated with) an outcome can lead to more extreme implicit evaluations; however, it remains unclear whether spontaneous mental state inferences contributed to this effect and whether inferences about intent in the absence of observable outcomes are sufficient for implicit evaluations to shift. Therefore, the present work was designed to provide initial evidence on the separate and combined effects of outcomes (directly observable consequences in the external world) and intent (indirectly inferred internal states of moral agents) on implicit evaluations. In doing so, our main goal is to further elucidate the fundamental nature of the computations underlying implicit evaluation. In addition, the present results may also enhance understanding of real-world processes of decision making in situations with moral implications. For instance, knowing how implicit evaluations of actors are guided by outcome and intent may provide some insight into everyday intuitions of blameworthiness.

Why might implicit (indirectly measured) evaluations behave differently from their explicit (directly measured) counterparts in response to the same morally relevant variables of intent and outcome? Explicit and implicit evaluations obviously differ from each other in features of the measurement context: Explicit evaluations are elicited via self-report, whereas implicit evaluations are elicited using less direct

measures that do not require intentional evaluative judgments on the participant's part. In addition to these differences in methods of measurement, most dual-process theories of social cognition (e.g., Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004) make the crucial claim that explicit and implicit evaluations emerge from fundamentally different processes of learning and representation. Specifically, these theories posit that whereas explicit evaluations are subserved by high-level inferential reasoning, implicit evaluations merely track co-occurrences of targets with valenced stimuli.

If this is, indeed, the case, then it seems reasonable to assume that implicit evaluations will be selectively sensitive to outcomes but not to intent. After all, a precondition for any causal relationship is closeness of the cause to the outcome in space and time (in this case, the moral agent to an event with moral implications, such as harm to another person). As such, if a moral agent becomes causally responsible for a valenced outcome (such as someone getting killed), then they should also become associated with the valence of that outcome simply by virtue of spatiotemporal proximity. For instance, in the story described above, a person letting someone fall to their death crossing an unstable bridge, even in the absence of any intention of doing so, may become linked with negative valence as a result of relatively low-level processes of association formation.[1]

However, the same reasoning does not apply to intent. For example, a person may try to unsuccessfully kill someone by letting the victim cross a bridge they erroneously believe to be unstable. In the absence of a manifest negative outcome, an observer can assign blame to this moral agent only if the observer (*a*) represents the agent's false belief about the state of the world and (*b*) encodes the agent's behavior of intentionally not warning the other person about the condition of the bridge. Without these preconditions being met, the presence of malicious intent cannot be inferred. Accordingly, if, as asserted by most dual-process theories, implicit evaluations are impervious to high-level inferential reasoning, they should selectively reflect outcomes but not intent.

In line with this logic, intent-based assignment of blame for moral actions seems to emerge later in development than outcome-based assignment of blame (Cushman et al., 2013; Zelazo, Helwig, & Lau, 1996), thus suggesting that the former may be subserved by more complex mental operations. Moreover, in a similar vein, placing participants under heavy cognitive load (Buon, Jacob, Loissel, & Dupoux, 2013; Martin, Buon, & Cushman, 2019) or interfering with neural activity related to mental state attribution (Young et al., 2010) appears to selectively disrupt intent-based moral reasoning. Specifically, such manipulations make attempted harms more permissible, while leaving outcome-based moral reasoning relatively intact.

By contrast, a more recent group of propositional theories (e.g., De Houwer, 2014; De Houwer, Van Dessel, & Moran, in press; Mitchell, De Houwer, & Lovibond, 2009) suggest that implicit evaluations of moral agents should be sensitive to much the same variables as their explicit counterparts. Indeed, in line with the propositional perspective, a growing body of empirical work has demonstrated the sensitivity of implicit evaluations to input that had traditionally been assumed to affect only explicit, but not implicit, cognition (for reviews see Cone, Mann, & Ferguson, 2017; De Houwer et al., in press). For instance, implicit evaluations have been shown to encode content far exceeding mere stimulus associations, such as relational qualifiers (Zanon, De Houwer, Gast, & Smith, 2014), reasoning about believability (Cone, Flaharty, & Ferguson, 2019), and correct and erroneous propositional inferences (Kurdi & Dunham, 2019).

Under propositional accounts, both explicit and implicit evaluations are posited to emerge from the same processes of propositional learning and representation, with the only major difference between the two consisting in the degree of automaticity with which propositions are retrieved from long-term memory. At the same time, propositional accounts can explain dissociations between explicit and implicit cognition by appealing to the incomplete retrieval of propositions at test rather than the presence of separate associative and propositional representations in memory (Van Dessel, Gawronski, & De Houwer, 2019). Given that the conditions under which such incomplete retrieval is posited to occur are currently not adequately specified, research on the types of propositional reasoning that can contribute to implicit evaluation constitutes an important step toward designing falsifiable propositional models of implicit social cognition.

Finally, in view of the extensive empirical evidence discussed above, some dual-process theories of social cognition now recognize the possibility that propositional inferences may, at least some of the time, indirectly influence implicit evaluation (Gawronski & Bodenhausen, 2006, 2011). As such, we do not believe that the present work should be seen as conclusively arbitrating between two broad classes of currently available theories. Rather, it is our hope that the evidence provided here will have the ability to constrain any current or future account of the basic computations from which implicit social cognition emerges.

2. Study 1

Previous work has found that negative outcomes and negative intent can each be sufficient to engender negative evaluations of moral agents when explicit (self-report) measures are used to index those evaluations (Cushman, 2008; Cushman et al., 2013; Knobe, 2005; Young et al., 2007; Young et al., 2010). At the same time, it is an open question whether implicit (indirectly measured) evaluations are also similarly modulated by outcome and intent. In this study, we sought to provide some initial evidence on this issue by assessing implicit evaluations of two moral agents: a moral agent who performed an action with a positive outcome and positive intent (control person) and another moral agent who performed an action with varying outcomes and intentions (target person).

An effect of outcome on implicit evaluation could reasonably be anticipated within either a dual-process (Gawronski & Bodenhausen, 2006, 2011; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004) or a single-process propositional (De Houwer, 2014; De Houwer et al., in press; Mitchell et al., 2009) framework. However, if implicit cognition is to adaptively guide social behavior, negative intent should also be sufficient to produce negative implicit evaluations of an individual: A person with negative intent can be expected to try performing harmful actions in the future even if the first attempt was not successful. Nonetheless, under most dual-process theories of social cognition, implicit evaluations are posited to reflect only associative information but not high-level reasoning about an actor's unobservable mental states (but see Gawronski & Bodenhausen, 2006, 2011). In contrast, the more recent single-process propositional perspective would be able to accommodate an effect of high-level reasoning about intent on implicit evaluations.

2.1. Method

We report all measures, manipulations, and exclusions in both studies.

2.1.1. Participants and design

Participants were 624 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu). The sample size was determined before any data analysis.

[1] Previous findings by Cone and Ferguson (2015), Hughes et al. (2019), and Kurdi et al. (2020) suggest that a causal relationship of this kind cannot be reduced to a pure association. In other words, a person who is causally responsible for an outcome (e.g., someone getting killed) should be subject to more negative implicit evaluations than someone who was merely present when the same outcome occurred. However, this distinction is not the subject of the present investigation.

Table 1
Sample vignettes for Study 1. Participants read a vignette about a control person and a vignette about a target person in counterbalanced order. The content of the vignette about the target person was determined based on the participant's condition assignment. Names were randomly selected from a list of 10 for each vignette. Screen captures of the entire paradigm, as well as the text of all 12 vignettes, are available on the Open Science Framework (OSF; https://osf.io/nt596/).

Control person	Target person		
	Accident condition	Attempt condition	Harm condition
Jake and an acquaintance are camping in the woods. Jake spots some wild mushrooms growing along the campsite. Jake studies the mushrooms and consults his plant life guide. // The mushrooms happen to be edible and delicious. They are the kind that one can buy in the supermarket and put in salad. // Jake sees a picture of an edible mushroom in his book that looks just like these mushrooms at the campsite, so he believes that the mushrooms are edible. // Jake offers the mushrooms to his acquaintance. His acquaintance eats them and finds them very tasty.	Scott and his girlfriend are hiking. They come across a long narrow bridge that spans a steep canyon. // The bridge happens to be extremely unsteady and cannot carry the weight of even one very light person. // Scott believes that whoever walks on the bridge will cross the canyon quite safely because the bridge is maintained by the national park. // Scott says nothing as his girlfriend starts walking across the bridge. His girlfriend breaks the bridge and falls to her death.	Scott and his girlfriend are hiking. They come across a long narrow bridge that spans a steep canyon. // The bridge happens to be extremely sturdy and can easily carry the weight of many people at once. // Scott believes that whoever walks on the bridge will break the bridge and fall into the canyon because the bridge looks unsteady and old. // Scott says nothing as his girlfriend starts walking across the bridge. His girlfriend reaches the other side safely.	Scott and his girlfriend are hiking. They come across a long narrow bridge that spans a steep canyon. // The bridge happens to be extremely unsteady and cannot carry the weight of even one very light person. // Scott believes that whoever walks on the bridge will break the bridge and fall into the canyon because the bridge looks unsteady and old. // Scott says nothing as his girlfriend starts walking across the bridge. His girlfriend breaks the bridge and falls to her death.

Demographic variables, including age, gender, race, ethnicity, educational attainment, and political ideology, are available in the data files posted on OSF (https://osf.io/nt596/) for both studies, but were not analyzed for the purposes of the present project.

In line with standard practice, participants who did not complete the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which served as the main dependent measure (N = 17), and participants whose response latencies were below 300 ms on at least 10% of IAT trials (N = 7) were excluded from further analyses (Greenwald, Nosek, & Banaji, 2003). Moreover, in line with standard practice in learning studies of this kind (e.g., Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016; Van Dessel, Gawronski, Smith, & De Houwer, 2017) and to increase the internal validity of the study, participants who did not show perfect performance on four manipulation check items probing their explicit memory for outcomes and beliefs in the control and target vignettes (see below; N = 213) were also excluded from consideration. These participant exclusions resulted in a final sample size of 387. Participants were randomly assigned to one of three conditions: accident (N = 145), attempt (N = 123), and harm (N = 119).

Based on a sensitivity power analysis, this sample size provides 0.80 power for the detection of an effect size of f^2 = 0.025, which is smaller than the effect size obtained in both critical regression models (f^2 = 0.238 in the regression model for explicit evaluations and f^2 = 0.033 in the regression model for implicit evaluations). That is, the design was adequately powered to detect even considerably smaller effects than the ones that were obtained in the actual study.

2.1.2. Materials

2.1.2.1. Vignettes. 12 vignettes were adapted from Young et al. (2007) for use in the present experiment, with small modifications necessitated by differences in the design of the two studies. All vignettes (for examples, see Table 1) describe a protagonist performing an act with moral implications. Each vignette was created in four versions, with outcome (positive vs. negative) and intent (positive vs. negative) manipulated orthogonally. For instance, in one of the vignettes, the protagonist's girlfriend is about to cross a bridge on a hike. In the positive outcome version of the vignette, the bridge is steady and the girlfriend is unharmed, whereas in the negative outcome version, the bridge is old and dangerous and the girlfriend falls to her death. In the positive intent version, the protagonist believes that the bridge is stable, whereas in the negative intent version, the protagonist believes that the bridge is unstable. Accordingly, crossing these two factors yields four different versions of each vignette: harmless (positive intent + positive outcome), accident (positive intent + negative outcome), attempt (negative intent + positive outcome), and harm (negative intent + negative outcome).

2.1.2.2. Names. Ten men's names, all widely used among White Americans, were selected for use as names for the control and target individuals (see below). The names included Brett, Cody, Dustin, Garrett, Jake, Luke, Max, Scott, Tanner, and Wyatt.

2.1.2.3. Faces. Images depicting the faces of four White men, drawn from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015), were selected for use as control and target individuals. For each of the four individuals, all four versions of the face, including neutral, happy (with closed mouth), happy (with open mouth and visible teeth), and angry, were included in the study.[2] The ID of each image is available on OSF (https://osf.io/nt596/). In line with the instructions accompanying the Chicago Face Database, the images can be obtained directly from https://chicagofaces.org/default/.

2.1.2.4. Trait adjectives. The trait adjectives “honest,” “honorable,” “moral,” “sincere,” and “trustworthy” were selected for use as positive attributes and the trait adjectives “cruel,” “deceitful,” “dishonorable,” “malicious,” and “mean” for use as negative attributes on the implicit and explicit evaluation tasks.

2.1.3. Procedure and measures

The study consisted of a learning phase and a test phase. In the learning phase, participants read two vignettes. One of these vignettes described a control person who had positive intent and performed an action with a positive outcome. The other vignette described a target person who (*a*) had positive intent but performed an action with a negative outcome (accident condition), (*b*) had negative intent but performed an action with a positive outcome (attempt condition), or (*c*)

[2] These four versions of the faces were included not for a theoretical reason but rather because the IAT requires multiple unique stimuli for each category. We did not find any evidence that the emotional content conveyed by the faces interfered with the measurement of implicit evaluations. A two-way ANOVA with critical block (congruent vs. incongruent), emotional content (neutral, happy/closed mouth, happy/open mouth, and angry), and their interaction as predictors and response latency as the dependent variable yielded only a main effect for critical block, $F(1, 23{,}672) = 69.54$, $p < .001$. The main effect of emotional content, $F(3, 23{,}672) = 1.83$, $p = .139$, and the interaction, $F(3, 23{,}672) = 0.30$, $p = .829$, were not statistically significant. Moreover, in a Bayesian model comparison framework, the posterior probability of the linear model including only a main effect for critical block was $p > .999$.

had negative intent and performed an action with a negative outcome (harm condition). In the test phase, participants completed an Implicit Association Test (IAT; Greenwald et al., 1998) measuring implicit evaluations of the control person relative to the target person, followed by a set of explicit evaluation items and a set of items probing explicit memory for crucial details of the two vignettes. A screen capture of the entire procedure, as well as the verbatim text of the initial instructions, vignettes, and explicit items, is available for download from OSF (https://osf.io/nt596/) for both studies.

2.1.3.1. Learning phase. At the outset of the learning phase, participants were introduced to two individuals, along with their names and faces, and were told that they would learn about these two individuals. The names and faces of the two individuals were randomly drawn from the names and faces described above. Participants were then informed that they would read a story about each individual and were asked to form an impression of them. Impression formation instructions have been shown to boost learning in similar tasks (e.g., Moran et al., 2015).

Following the initial instructions, participants were exposed to two vignettes, one about each individual. Each of the two vignettes was randomly drawn from the 12 vignettes described above, with the constraint that the same participant was never exposed to different versions of the same vignette. For instance, if a participant received the mushroom vignette (see Table 1) as the control vignette, the target vignette was drawn from the remaining 11 vignettes under exclusion of the mushroom vignette. Each vignette was presented over four screens, with the face of the control or target person, as applicable, placed above the text on each screen to facilitate learning. The order of the two vignettes was counterbalanced.

The vignette about the control person described a story involving a moral act with positive intent and a positive outcome. For instance, Jake may offer his acquaintance some mushrooms believing that they are edible (positive intent); in fact, the mushrooms turn out to be delicious and his acquaintance enjoys them (positive outcome). The crucial features of the vignette about the target person depended on the participant's condition assignment. In the accident condition, the story involved a moral act with positive intent and a negative outcome. For instance, Scott may let his girlfriend cross the bridge believing that it is stable (positive intent); in fact, the bridge is unstable and the girlfriend dies (negative outcome). In the attempt condition, the story involved a moral act with negative intent and a positive outcome. For instance, Scott may let his girlfriend cross the bridge believing that it is unstable (negative intent); in fact, the bridge is stable and the girlfriend is fine (positive outcome). Finally, in the harm condition, the story involved a moral act with both negative intent and a negative outcome. For instance, Scott may let his girlfriend cross the bridge believing that it is unstable (negative intent); in fact, the bridge is unstable and the girlfriend dies (negative outcome). As such, this design allowed us to probe the separate and joint effects of intent and outcome on implicit evaluations.

2.1.3.2. Test phase. In the test phase, participants completed an IAT (Greenwald et al., 1998) measuring implicit evaluations of the control person relative to the target person. The IAT was followed by (*a*) a battery of explicit evaluation items that used the same stimuli as the IAT and then (*b*) four explicit memory items probing participants' recollection for crucial details of the learning phase. Given that the present project focuses primarily on implicit evaluations, the IAT was always administered before the explicit measures.

2.1.3.2.1. Implicit evaluations. Implicit evaluations of the control person relative to the target person were measured using a standard five-block IAT (Greenwald et al., 1998). The IAT is a response interference task on which implicit evaluations are inferred by comparing participants' speed and accuracy across two blocks of combined sorting trials: a first combined block in which one target (e.g., the control person) shares a response key with positive items and the other target (e.g., the target person) shares a response key with negative items, and a second combined block in which the assignment of targets to valences is reversed (e.g., target person–positive vs. control person–negative).

In block 1 (category practice block; 20 trials), participants used two response keys (E and I) to sort the four face images associated with the control person and target person each. The names previously assigned to the control person and the target person (e.g., “Jake” and “Scott”) served as category labels. In block 2 (attribute practice block; 20 trials), participants sorted the positive and negative trait adjectives described above. The words “good” and “bad” served as attribute labels. In block 3 (first combined block; 40 trials), participants used one response key to sort images of the control person and positive trait adjectives and the other response key to sort images of the target person and negative trait adjectives. In block 4 (reversed category practice block; 20 trials), participants sorted the images of the control person and the target person used in blocks 1 and 3 but with the mapping of categories to response keys reversed. Finally, in block 5 (second combined block; 40 trials), participants used one response key to sort images of the target person and positive trait adjectives and the other response key to sort images of the control person and negative trait adjectives.

Although the above description of the IAT contains the congruent (i.e., control person/positive–target person/negative) combined block administered first, in fact, the order of combined blocks was counterbalanced. Performance on the IAT was assessed using the improved scoring algorithm (Greenwald et al., 2003) such that higher D scores index relatively more positive evaluations of the control person and relatively more negative evaluations of the target person, in line with the theoretically expected direction of the effect. The IAT showed adequate internal consistency (r = 0.73 in Study 1 and r = 0.68 in Study 2 based on 500 split-half correlations).

2.1.3.2.2. Explicit evaluations. Following the IAT, participants were asked to indicate to what extent they thought each of the five positive trait adjectives and each of the five negative trait adjectives used on the IAT characterized the control person and the target person, respectively. We used the same attribute stimuli on the implicit and explicit measures of evaluation to minimize differences between the two measures that are unrelated to differences in the cognitive processes that they are designed to capture (Gawronski, 2019). The same explicit items were administered on separate pages for the control person and for the target person. The order of the two pages was counterbalanced. Within each page, the order of trait adjectives was individually randomized, with positive and negative trait adjectives intermixed. To remind participants of the identity of the control and the target person, their names were included in the questions and their pictures were shown on top of the page. For each item, the response options ranged from 1 (labeled “not at all characteristic”) to 7 (labeled “extremely characteristic”).

Negative items were reverse scored to ensure that higher scores on all explicit items corresponded to more positive evaluations. Explicit evaluations of the control person (Cronbach's α = 0.93 in Study 1 and Cronbach's α = 0.89 in Study 2) and of the target person (Cronbach's α = 0.95 in Study 1 and Cronbach's α = 0.94 in Study 2) were highly reliable. Therefore, they were averaged to form a single index of explicit evaluation for the control person and the target person each. Finally, to make explicit evaluation scores comparable to implicit evaluation scores, explicit evaluations of the target person were subtracted from explicit evaluations of the control person.

2.1.3.2.3. Explicit memory. The present study cannot form the basis of valid inferences about the effects of outcome and intent on implicit evaluations unless participants have the ability to correctly infer both from the vignettes that they have read. Therefore, participants were asked to report, both with regard to the control person and the target person, whether (*a*) anything bad happened to the other person in the story (outcome) and (*b*) whether the control or target person believed

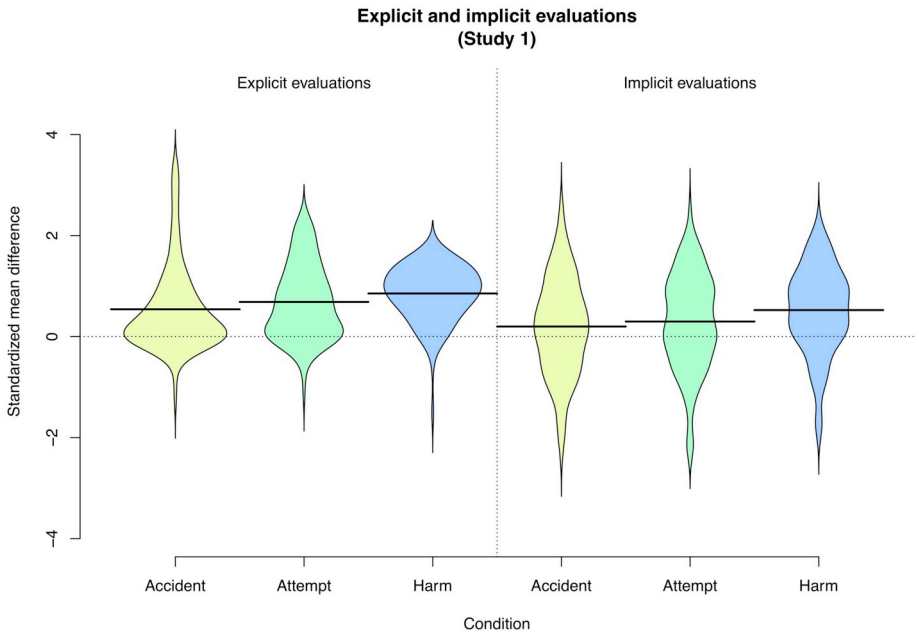

Fig. 1. Distribution of explicit and implicit evaluations by condition (Study 1), displayed in standardized units to ensure comparability. The dashed horizontal line shows neutrality and the solid horizontal lines show the means of the explicit and implicit measures comparing the target person to the control person. Positive scores indicate the theoretically expected preference for the control person over the target person.

that anything bad would happen (irrespective of whether it actually happened; intent). Whether the two items were administered with respect to the control person or the target person first was counterbalanced. Within each person, the outcome item was always administered first and the intent (belief) item was always administered second. Similar to the explicit evaluation items, the corresponding names and faces were included on each page to aid recollection of the identity of the two individuals. As described above, participants who did not perform perfectly on the explicit memory items were excluded from further analyses. However, their data are available for reanalysis to interested researchers from OSF (https://osf.io/nt596/).

2.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 1. Descriptively, both measures seemed to exhibit a preference in favor of the control person (who performed an action with positive intent and a positive outcome) over the target person, as reflected by positive average explicit evaluations in the accident (M = 0.98, SD = 1.53), attempt (M = 1.87, SD = 1.98), and harm conditions (M = 3.44, SD = 2.08), as well as positive average implicit evaluations in the accident (M = 0.09, SD = 0.45), attempt (M = 0.14, SD = 0.46), and harm conditions (M = 0.29, SD = 0.47).

2.2.1. Explicit evaluations

Explicit evaluations of the control person relative to the target person were examined using a linear model with experimental condition as the sole predictor and the accident condition as the reference category.[3] The overall model was found to be statistically significant, F (2, 380) = 56.96, p < .001, R^2 = 0.23, suggesting differences in explicit evaluation across the three conditions. The intercept was positive and statistically significant, b = 0.98, 95% CI: [0.68; 1.29], t (380) = 6.37, p < .001, indicating that the control person was preferred to the target person in the accident condition. The slope for the attempt condition was also positive and statistically significant, b = 0.88, 95% CI: [0.43; 1.33], t(380) = 3.86, p < .001, indicating that the preference for the control person over the target person was stronger in the attempt than in the accident condition. Finally, the slope for the harm condition was also positive and statistically significant, b = 2.45, 95% CI: [2.00; 2.90], t(380) = 10.62, p < .001, indicating that the preference for the control person over the target person was stronger in the harm than in the accident condition.

However, the model discussed above does not explicitly account for potential effects of variation in stimuli, including vignettes, names, and faces. Therefore, we also fit a Bayesian mixed-effects model[4] containing a fixed effect for experimental conditions and random intercepts for control vignette–target vignette pairs, control name–target name pairs, and control face–target face pairs. The results that emerged were inferentially identical to the results obtained using the more parsimonious linear model. Specifically, a preference for the control person over the target person was revealed in the accident condition, b = 0.99, 95% HDI: [0.73; 1.25]. This preference was found to be stronger in the attempt condition, b = 0.87, 95% HDI: [0.50; 1.25], and even stronger in the harm condition, b = 2.44, 95% HDI: [2.06; 2.82].

2.2.2. Implicit evaluations

Implicit evaluations of the control person relative to the target person were also examined using a linear model with experimental condition as the sole predictor and the accident condition as the reference category. The overall model was found to be statistically significant, F(2, 384) = 6.49, p = .002, R^2 = 0.03, suggesting differences in implicit evaluation across the three conditions. The intercept was positive and statistically significant, b = 0.09, 95% CI: [0.02; 0.17], t (384) = 2.39, p = .017, indicating that the control person was preferred to the target person in the accident condition. Unlike for explicit evaluations, the slope for the attempt condition was not statistically significant, b = 0.05, 95% CI: [−0.06; 0.16], t(384) = 0.93, p = .352, indicating that the extent of preference for the control person over the target person was similar across the attempt and accident conditions. Finally, the slope for the harm condition was positive and statistically significant, b = 0.20, 95% CI: [0.09; 0.31], t(384) = 3.52, p < .001, indicating that the preference for the control person over the target person was stronger in the harm than in the accident condition.

Inferentially identical results were obtained in a Bayesian mixed-

[3] The accident condition was used as the reference category because previous research has found the smallest effect in this condition (e.g., Young et al., 2007).

[4] We opted for the Bayesian modeling strategy because trying to fit the same models in a frequentist framework resulted in singularities (Gelman & Hill, 2006).

effects model containing a fixed effect for experimental conditions and random intercepts for control vignette–target vignette pairs, control name–target name pairs, and control face–target face pairs. Specifically, a preference for the control person over the target person was revealed in the accident condition, $b = 0.09$, 95% HDI: [0.01; 0.17], with a comparable preference emerging in the attempt condition, $b = 0.05$, 95% HDI: [−0.05; 0.14]. Finally, the preference was found to be stronger in the harm condition than in the accident condition, $b = 0.19$, 95% HDI: [0.10; 0.29].

2.3. Discussion

Study 1 replicated previous results obtained using explicit measures (e.g., Young et al., 2007): Negative intent and negative outcome were each found to be sufficient to produce negative evaluations of a moral agent relative to another moral agent whose actions reflected positive intent and resulted in a positive outcome. Crucially, we also newly demonstrated that implicit evaluations of the same moral agents, as measured using an Implicit Association Test, can show similar trends, with significant differences between the control person and target person emerging in all three experimental conditions. As such, these findings suggest that although observable negative outcomes can produce negative implicit evaluations of moral agents, they are not necessary for negative implicit evaluations to emerge. Rather, inferences about a person's directly unobservable intentions, and presumably their likely future behavior in light of those intentions, seem to be sufficient. Moreover, also in line with an inferential account, we found implicit evaluations resulting from the same observable adverse outcome to be more negative if the agent was described as acting with malicious rather than benign intentions.

3. Study 2

The results of Study 1 suggest that implicit evaluations of moral agents can reflect negative outcomes in the absence of negative intent and, critically, negative intent in the absence of negative outcomes. Given the novelty of this finding, we conducted an additional test of the same idea using a slightly different, and more conservative, design. Specifically, unlike in Study 1, where the control person was described as performing a moral act with positive intent and outcome, all control vignettes in the present study described a person performing a morally neutral act. This neutral baseline provides a more exacting reference point for the effects of negative outcome and intent to emerge. Moreover, negative evaluations in the accident and attempt conditions of the previous study may, at least in part, have emerged due to the presence of a false belief rather than, as intended, the moral implications of the protagonist's actions. As such, in the present study, control and target vignettes were matched with each other for the presence of a true or false belief.

In addition, to probe the generalizability of the findings from Study 1 to different stimuli, a new set of names and a new set of faces were used. Although the emotional content conveyed by the faces did not seem to interfere with the measurement of implicit evaluations in Study 1, in the present study we used only faces with neutral expressions. Finally, unlike in Study 1, the study design and analytic plan, including exclusion criteria, were preregistered (https://aspredicted.org/blind.php?x=bb4ns6). All analyses not included in the preregistration document are explicitly noted as exploratory below.

3.1. Method

3.1.1. Participants and design

The method of recruitment and exclusion criteria were identical to those used in Study 1. The sample size was determined before any data analysis and preregistered. Of an initial sample of 852, 24 participants were excluded for failing to complete the IAT (Greenwald et al., 1998) and 16 for excessively fast responding. Moreover, participants who did not show perfect performance on four manipulation check items probing their explicit memory for outcomes and beliefs in the control and target vignettes (see below; $N = 373$) were also excluded from consideration. These participant exclusions resulted in a final sample size of 439. Whereas Study 1 included three conditions (accident, attempt, and outcome), each compared to a harmless baseline, participants in the present study were randomly assigned to one of four conditions: harmless (positive intent + positive outcome; $N = 134$), accident (positive intent + negative outcome; $N = 120$), attempt (negative intent + positive outcome; $N = 78$), and harm (negative intent + negative outcome; $N = 107$), each compared to a morally neutral baseline.

Based on a sensitivity power analysis, this sample size provides 0.80 power for the detection of an effect size of $f^2 = 0.025$, which is smaller than the effect size obtained in both critical regression models ($f^2 = 0.264$ in the regression model for explicit evaluations and $f^2 = 0.039$ in the regression model for implicit evaluations). That is, the design was adequately powered to detect even considerably smaller effects than the ones that were obtained in the actual study.

3.1.2. Vignettes

12 target vignettes and 12 control vignettes were adapted from Chakroff et al. (2015) for use in the present study, with small modifications necessitated by differences in experimental designs. Target vignettes were similar to the ones used in Study 1, with all of them describing a protagonist performing an act with moral implications (for examples, see Table 2). Each vignette was created in four versions, with outcome (positive vs. negative) and intent (positive vs. negative) manipulated orthogonally. Unlike in Study 1, the control vignettes did not include actions with moral implications. Rather, they described stories involving neutral actions and either a true belief or a false belief. For instance, the protagonist may observe a big toad during his walk in the park. He may (true belief) or may not (false belief) have known that the area had toads. In a further deviation from Study 1, vignettes described outcomes first and beliefs second.

3.1.3. Names

Ten new men's names, all widely used among White Americans, were selected for use as names for the control and target individuals (see below). The names included Anthony, Christopher, David, Ethan, Henry, James, Lucas, Michael, Oliver, and Ryan.

3.1.4. Faces

Images depicting the faces of eight White men were selected for use as control and target individuals from the Radboud Faces Database (Langner et al., 2010). For each individual, three versions of the face, differing in camera angle (45 degrees to the left, frontal, and 45 degrees to the right), were included in the study. Unlike in Study 1, all faces used had neutral emotional expressions. The ID of each image is available on OSF (https://osf.io/nt596/). In line with the instructions accompanying the Radboud Faces Database, the images can be obtained directly from http://www.socsci.ru.nl:8180/RaFD2/RaFD.

3.1.5. Trait adjectives

The same trait adjectives were used as in Study 1.

3.1.6. Procedure and measures

The procedure and measures were highly similar to Study 1: Participants read two vignettes, one about a control person and one about a target person, followed by an IAT (Greenwald et al., 1998) measuring implicit evaluations of the two individuals, and finally by a battery of Likert items probing explicit evaluations and explicit memory for crucial details of the learning phase.

However, some important changes were implemented. Specifically, in Study 1, a harmless (positive intent + positive outcome) vignette

Table 2
Sample vignettes for Study 2. Participants read a vignette about a control person and a target person in counterbalanced order. The content of the vignette about the target person was determined based on the participant's condition assignment. The vignette about the control person involved a true belief in the harmless and harm conditions and a false belief in the accident and attempt conditions. Names were randomly selected from a list of 10 for each vignette. Screen captures of the entire paradigm, as well as the text of the 12 control and 12 target vignettes, are available on the Open Science Framework (OSF; https://osf.io/nt596/).

Control person	Target person			
True belief	False belief	Harmless condition	Accident condition	Attempt condition	Harm condition
James is taking a stroll in the park by his house along the bank of a stream. It is finally nice weather outside, and James is enjoying the fresh cool air. // Every once in a while, James picks up a stone and skips it across the water. // As he reaches for a stone, a big toad suddenly hops from next to it into the water. // James knew that the area had toads.	James is taking a stroll in the park by his house along the bank of a stream. It is finally nice weather outside, and James is enjoying the fresh cool air. // Every once in a while, James picks up a stone and skips it across the water. // As he reaches for a stone, a big toad suddenly hops from next to it into the water. // James didn't know that the area had toads.	Michael is grocery shopping for his grandmother. Bagged spinach had recently been recalled for *E. coli* contamination, but some supermarkets have begun carrying it again. // Michael buys spinach for his grandmother and uses it to make her a large salad. // The spinach is perfectly safe to eat and Michael's grandmother will enjoy the salad very much. // Michael had checked online and knew that the spinach at the supermarket was not contaminated.	Michael is grocery shopping for his grandmother. Bagged spinach had recently been recalled for *E. coli* contamination, but some supermarkets have begun carrying it again. // Michael buys spinach for his grandmother and uses it to make her a large salad. // The spinach is contaminated with *E. coli* and will make Michael's grandmother very sick. // Michael had checked online and believed that the spinach at the supermarket was contaminated.	Michael is grocery shopping for his grandmother. Bagged spinach had recently been recalled for *E. coli* contamination, but some supermarkets have begun carrying it again. // Michael buys spinach for his grandmother and uses it to make her a large salad. // The spinach is perfectly safe to eat and Michael's grandmother will enjoy the salad very much. // Michael had checked online and believed that the spinach at the supermarket was contaminated.	Michael is grocery shopping for his grandmother. Bagged spinach had recently been recalled for *E. coli* contamination, but some supermarkets have begun carrying it again. // Michael buys spinach for his grandmother and uses it to make her a large salad. // The spinach is contaminated with *E. coli* and will make Michael's grandmother very sick. // Michael had checked online and knew that the spinach at the supermarket was contaminated.

always served as the control vignette, with the remaining three combinations of intent and outcome implemented in target vignettes. By contrast, in the present study, target vignettes reflected orthogonal between-participant manipulations of intent (positive vs. negative) and outcome (positive vs. negative), including the harmless (positive intent + positive outcome) case. As such, we were able to explicitly model the separate and joint effects of these two variables.

Moreover, unlike in Study 1, a vignette involving no moral acts but either a true or a false belief served as a control vignette. Because the harmless (positive intent + positive outcome) and harm (negative intent + negative outcome) conditions involved true beliefs, the control vignette in these conditions also involved a true belief. By contrast, the accident (positive intent + negative outcome) and attempt (negative intent + positive outcome) conditions involved false beliefs; as such, the control vignette in these conditions also involved a false belief. Finally, given that control and target vignettes were drawn from separate pools, any control vignette could be paired with any target vignette, without restrictions.

3.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 2. As expected, in the harmless condition, evaluations of the control person and the target person were similar to each other on both explicit (M = 0.20, SD = 1.31) and implicit (M = 0.03, SD = 0.41) measures. In the remaining conditions, similar to Study 1, both measures seemed to exhibit a preference in favor of the control person over the target person, as reflected by positive average explicit evaluations in the accident (M = 0.67, SD = 1.52), attempt (M = 0.49, SD = 1.61), and harm conditions (M = 2.51, SD = 1.81), as well as positive average implicit evaluations in the accident (M = 0.17, SD = 0.44), attempt (M = 0.14, SD = 0.42), and harm conditions (M = 0.25, SD = 0.44).

3.2.1. Explicit evaluations

The linear model with explicit evaluations as the dependent variable and main effects for intent (positive vs. negative), outcome (positive vs. negative), and their interaction as predictors was found to be statistically significant, $F(3, 432) = 49.29$, $p < .001$, $R^2 = 0.26$, suggesting differences in explicit evaluation across the four conditions. The intercept (positive intent + positive outcome) was not statistically significant, $b = 0.21$, 95% CI: [−0.06; 0.47], $t(432) = 1.54$, $p = .124$, indicating that explicit evaluations of the target person were similar to explicit evaluations of the control person in the harmless condition. The slope for outcome was positive and statistically significant, $b = 0.47$, 95% CI: [0.08; 0.85], $t(432) = 2.39$, $p = .017$, indicating that a negative outcome increased the difference between the control and target person in the theoretically expected direction. Unexpectedly, although positive, the slope for intent was not statistically significant, $b = 0.29$, 95% CI: [−0.15; 0.72], $t(432) = 1.29$, $p = .196$, indicating that negative intent, in and of itself, was not sufficient to reliably increase the difference in explicit evaluation between the control person and the target person. Finally, the slope for the outcome × intent interaction was positive and statistically significant, $b = 1.55$, 95% CI: [0.95; 2.14], $t(432) = 5.09$, $p < .001$, indicating that the joint effects of negative outcome and negative intent exceeded the separate effects of the two variables.

Inferentially identical results were obtained in a Bayesian mixed-effects model containing the same fixed effects as the linear model discussed above, along with random intercepts for control vignette–target vignette pairs, control name–target name pairs, and control face–target face pairs. Specifically, we found no significant difference between the control person and target person in the positive outcome + positive intent (harmless) condition, $b = 0.20$, 95% HDI: [−0.03; 0.43], a significant effect for outcome, $b = 0.47$, 95% HDI: [0.16; 0.80], no significant effect for intent, $b = 0.32$, 95% HDI: [−0.05; 0.70], and a significant outcome × intent interaction,

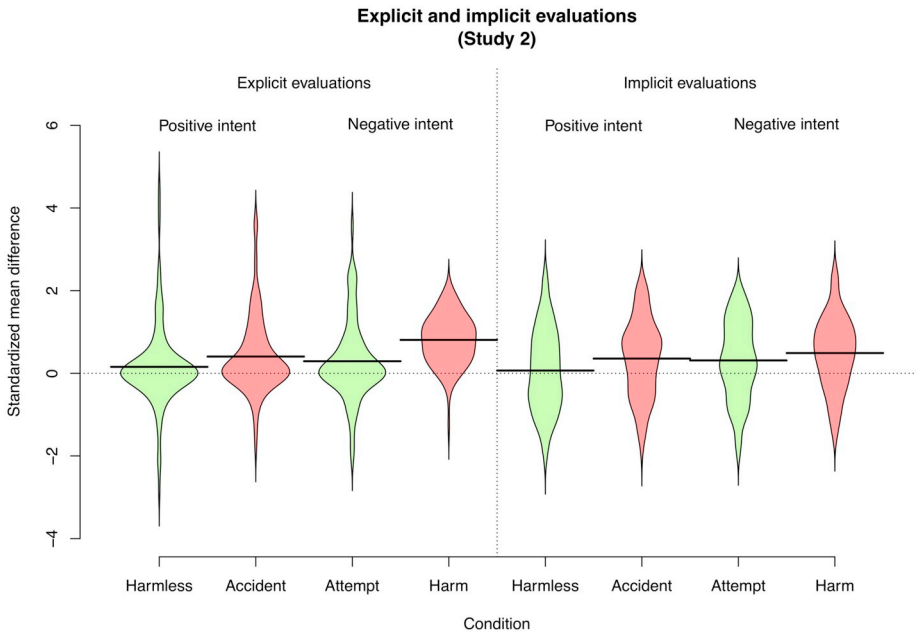

Fig. 2. Distribution of explicit and implicit evaluations by condition (Study 2), displayed in standardized units to ensure comparability. The dashed horizontal line shows neutrality and the solid horizontal lines show the means of the explicit and implicit measures comparing the target person to the control person. Positive scores indicate the theoretically expected preference for the control person over the target person. Conditions with positive outcome are shown in green and conditions with negative outcome are shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

b = 1.50, 95% HDI: [1.00; 2.02].

In a set of exploratory (non-preregistered) analyses, we sought to investigate why, unexpectedly, the main effect of intent was not statistically significant in either model. We found that the control person was evaluated more negatively in the conditions involving false beliefs than in the conditions involving true beliefs, b = 0.29, 95% CI: [−0.43; −0.05], $t(435)$ = 2.45, p = .014. This result suggests the presence of a halo effect (Nisbett & Wilson, 1977): Simply by virtue of possessing a morally irrelevant false belief, the control person was evaluated as relatively less morally virtuous. Moreover, the effect of intent was found to be significant when explicit evaluations of the target person, rather than relative explicit evaluations of the target person compared to the control person, were used as the dependent measure in a linear model, b = 0.54, 95% CI: [0.20; 0.87], $t(432)$ = 3.16, p = .002. To summarize, a halo effect with regard to the control person may have unexpectedly attenuated the effect of the manipulation of intent with regard to the target person when the relative evaluation variable was used as the dependent measure in the preregistered models. However, given that the absolute measure of evaluation produced the theoretically expected result and, as preregistered, the main focus of the present studies is on implicit, rather than explicit, social cognition, we feel sufficiently confident to proceed with the interpretation of the results obtained on the implicit measure of evaluation.

3.2.2. Implicit evaluations

The linear model with implicit evaluations as the dependent variable and main effects for intent (positive vs. negative), outcome (positive vs. negative), and their interaction was found to be statistically significant, $F(3, 435)$ = 5.59, p < .001, R^2 = 0.03, suggesting differences in implicit evaluation across the four conditions. The intercept (positive intent + positive outcome) was not statistically significant, b = 0.03, 95% CI: [−0.04; 0.10], $t(435)$ = 0.74, p = .460, indicating that implicit evaluations of the target person were similar to implicit evaluations of the control person in the harmless condition. The slope for outcome was positive and statistically significant, b = 0.14, 95% CI: [0.03; 0.24], $t(435)$ = 2.60, p = .010, indicating that a negative, rather than positive, outcome increased the difference between the control and target person in the expected direction. The slope for intent was also positive and of similar magnitude, b = 0.11, 95% CI: [−0.01; 0.23], $t(435)$ = 1.85, p = .065, indicating that negative, rather than positive, intent increased the difference between the control person and the target person in the expected direction. However, this difference should be treated with some caution given that the p value accompanying it was just above α = 0.05. Finally, the slope for the outcome × intent interaction was not statistically significant, b = −0.03, 95% CI: [−0.19; 0.13], $t(435)$ = −0.36, p = .716, indicating that the joint effects of negative outcome and negative intent were not any stronger or weaker than would have been expected based on their separate effects.

Inferentially similar results were obtained in a Bayesian mixed-effects model containing the same fixed effects as the linear model discussed above, along with random intercepts for control vignette–target vignette pairs, control name–target name pairs, and control face–target face pairs, with one important deviation. Specifically, no significant difference was found between the control person and target person in the positive outcome + positive intent (harmless) condition, b = 0.03, 95% HDI: [−0.03; 0.09], a significant effect emerged for outcome, b = 0.14, 95% HDI: [0.05; 0.23], and no significant outcome × intent interaction was obtained, b = −0.03, 95% HDI: [−0.17; 0.10]. Crucially, unlike in the linear model above, the effect of intent was found to be unequivocally significant in the expected direction, b = 0.11, 95% HDI: [0.01; 0.21]. As such, this preregistered analysis suggests that not explicitly accounting for stimulus variation in the more parsimonious linear model above may, to some degree, have obscured the effect of intent.

3.3. Discussion

In Study 2, we replicated the focal finding of Study 1 using a different set of vignettes and target individuals and a more conservative (preregistered) design: Negative outcome and negative intent were each independently sufficient to produce negative implicit evaluations of moral agents relative to neutral controls whose actions did not carry moral implications. Moreover, the effect of both variables was found to be additive, with malicious rather than benign intent resulting in more negative implicit evaluations in the presence of identical observable outcomes.

4. General discussion

We conducted two high-powered experiments, one of them preregistered, to newly probe the separate and joint effects of outcomes

(observable consequences of actions in the external world) and intent (unobservable internal states) on the implicit evaluations of moral agents. The two studies yielded consistent evidence indicating that negative outcomes (e.g., causal responsibility for a person getting killed) and negative intent (e.g., a failed plan to kill a person) can each be independently sufficient for negative implicit evaluations to emerge. Moreover, in both studies, implicit evaluations originating from the same outcome were more negative if the moral agent acted with malicious, rather than benign, intent.

In Study 1, implicit evaluations were measured relative to a control person who had positive intent and whose actions resulted in a positive outcome. Study 2 had a more conservative design where implicit evaluations were compared to a control person who performed morally neutral actions but held beliefs whose veracity was matched to the veracity of the target person's beliefs. These findings generalized across variations in specific stimuli, including vignettes and target identities, as confirmed both by similar results across the two experiments and significant effects emerging in Bayesian mixed-effects models explicitly accounting for such stimulus variation.

These results extend a rapidly growing body of work indicating that implicit evaluations can flexibly respond to evaluative information, including information previously thought to influence only explicit cognition (Cone et al., 2017; De Houwer et al., in press). Specifically, the present studies provide evidence that implicit evaluations are sensitive not only to manifest outcomes (such as being causally responsible for a person falling to their death from an unstable bridge) but also to reasoning about an actor's unobservable mental states (such as plotting to kill someone by letting them cross an unstable bridge). Such responsiveness to negative intent in the absence of negative external consequences seems quite adaptive from the perspective of a view positing that the main function of social evaluation is the prediction of future social behavior (Tamir & Thornton, 2018), and especially the expected hedonic consequences of that future behavior for the self. A person who attempts to kill a friend once is likely to try killing the friend again and, as such, it seems warranted to attach a negative evaluation to them even if the first attempt does not succeed.

When it comes to currently available theories of implicit evaluation, these findings are more easily reconciled with recent propositional accounts (De Houwer, 2014; De Houwer et al., 2020; Mitchell et al., 2009) than most of their dual-process counterparts (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). The effect of directly observable outcomes on implicit evaluation can be accounted for within either a dual-process or a propositional framework. Specifically, dual-process theories may argue that a highly negative event (e.g., someone being causally responsible for another person's death) leads to the formation of an associative link between the responsible moral agent and negative valence. Propositional theories may explain the same effect by claiming that observers encode a causal relationship between the moral agent and the negative outcome and this causal relationship is then automatically activated on implicit measures of evaluation.

However, the effect of an actor's intentions is not as easily explained by dual-process theories. In particular, the effect of mere negative intent in the absence of a negative outcome, which in the present paradigm required reasoning about false beliefs to emerge, seems challenging to account for by a learning process relying purely on stimulus associations. At the same time, some dual-process theories leave some room for this kind of effect by positing that propositional reasoning can sometimes affect the conceptual associations underlying implicit evaluation (Gawronski & Bodenhausen, 2006, 2011). Nonetheless, even these theories postulate that implicit social cognition generally emerges from processes of association formation and we believe that accounting for the pattern of results obtained here does not require any assumption of purely associative processes.

In addition, the learning effects observed in any condition of the present studies may be seen as conflicting with certain other dual-process theories of social cognition (e.g., Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004) given that they show rapid updating of implicit evaluations in the face of a one-shot language-based intervention.[5] Indeed, these dual-process accounts posit that the updating of implicit evaluations should unfold in a slow and piecemeal manner rather than quickly and dynamically. At the same time, such exclusive reliance on gradual and incremental learning seems to uniquely characterize dual-process theories of social cognition rather than theories of associative learning more generally. For instance, the well-known and widely used Rescorla–Wagner model (Rescorla & Wagner, 1972) can account for rapid associative learning effects by positing a high learning rate. As such, dual-process theories of social cognition could also be quite easily modified to recognize this possibility, whereas accounting for the effects of inferential reasoning seems to require more fundamental changes.

At the same time, we do not see the present findings as in any way conflicting with previous work showing that intent-based reasoning is computationally complex, as evidenced by its relatively late emergence over the lifespan (Cushman et al., 2013; Zelazo et al., 1996) and its reliance on working memory (Buon et al., 2013; Martin et al., 2019). Specifically, the current results suggest that, once encoded, the outputs of intent-based moral judgment can be activated automatically. However, this need not imply that the process of arriving at this type of judgment is itself automatic (for a similar argument in the context of propositional reasoning see Kurdi & Dunham, 2019). In fact, making correct inferences about intent and outcome on the basis of the vignettes used in the present studies was challenging for participants: Although the modal performance in both studies was correct responding on both manipulation check items, a considerable number of participants made a mistake on either one (24.83% in Study 1; 31.77% in Study 2) or even both of them (10.67% in Study 1; 14.16% in Study 2).

Although the present studies principally sought to establish whether adverse outcomes (independent of the actor's intent) and malicious intent (independent of the action's outcome) are each independently sufficient for negative implicit evaluations to emerge, we also obtained data on the relative strength of both implicit and explicit evaluations across different learning conditions. In Study 1, malicious intent led to more negative evaluations than negative outcomes on explicit measures; on implicit measures, the two had comparable effects. Furthermore, whereas the joint effects of outcome and intent exceeded the separate effects of the two variables on both measures in Study 1 and on the explicit measure in Study 2, the interaction was not significant with implicit evaluations as the dependent measure in the latter study. Although these subtle differences were not our main focus in the present project, they make it clear that more research will be needed to explore the variations in and boundary conditions of the current findings.

Notably, both studies reported in the present paper used the Implicit Association Test (IAT; Greenwald et al., 1998) as their main dependent measure. Although the IAT and other implicit measures have been shown to behave similarly in both correlational (Bar-Anan & Vianello, 2018) and experimental designs (Bar-Anan & Nosek, 2017), some studies have also revealed differences. For instance, in recent work by Van Dessel, Ye, and De Houwer (2019), the IAT did not show sensitivity to novel information about well-known targets, whereas other implicit measures did. As such, it is conceivable that implicit measures other than the IAT would show patterns of updating different from the ones observed in the present project. We hope that this possibility will be investigated in future empirical work.

The present research may also be extended to other forms of moral reasoning not addressed in the current studies. In particular, all vignettes used here relied on harm-based moral violations, such as someone getting killed, injured, or psychologically hurt. However, some moral

[5] We thank an anonymous reviewer for bringing this point to our attention.

transgressions, such as consensual sibling incest or desecrating the American flag, constitute violations of purity without resulting in harm to another individual. Previous research has found that in the case of purity-based moral violations, outcomes tend to drown out the effects of intent (Chakroff et al., 2015; Haidt, 2001; Young & Saxe, 2011), thus suggesting that this distinction may modulate the findings obtained here. Additionally, the current studies focused on moral transgressions rather than morally virtuous behaviors: Even in the harmless (positive intent + positive outcome) conditions, participants were exposed to mildly positive behaviors with mildly positive outcomes, such as someone enjoying a delicious cup of coffee or a refreshing swim in a lake. These events are, of course, a far cry from the extremely positive events used in some previous work on the updating of implicit evaluations, such as someone donating a kidney to an unknown child (Cone & Ferguson, 2015).

Finally, all targets investigated in the present studies were young White men. Whiteness and maleness are seen as psychological defaults in present-day American society (Bosson, Vandello, & Buckner, 2018; Sue, 2004). As such, the use of White male targets may not have activated social group reasoning in most participants, thus allowing us to investigate the effects of intent and outcome in a relatively unconfounded manner. At the same time, we believe that the present paradigms could and should be extended to investigate the effects of moral reasoning on implicit evaluations in the context of other racial and gender groups.

Beyond considerations of equity and representativeness, there are specific reasons to believe that moral reasoning and the explicit and implicit evaluations resulting from it may differ as a function of target characteristics such as race and gender (Hester & Gray, 2020). For instance, research has shown that the same moral transgressions are evaluated differently if committed by ingroup vs. outgroup members (van der Toorn, Ellemers, & Doosje, 2015). Moreover, members of different racial and gender groups are subject to different stereotypes regarding their moral character (Fiske, Cuddy, Glick, & Xu, 2002; Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016). Finally, any effect of intent on implicit evaluations presupposes that the target activates mental state reasoning. However, previous research has shown that perceivers are less likely to mentalize about members of lower-status groups, especially extreme outgroups (Harris & Fiske, 2006; McLoughlin & Over, 2017). This suggests that moral reasoning, and explicit and implicit evaluations emanating from it, may be less sensitive to the intentions of outgroup targets relative to those of ingroup targets.

5. Summary

We conducted two well-powered experiments to demonstrate that implicit (indirectly measured) evaluations of moral actors can be sensitive not only to manifest outcomes (e.g., someone getting killed) but also to an actor's latent mental states (e.g., the intent to kill someone). This project strengthens connections between moral psychology, which overwhelmingly uses explicit measures of evaluation, and implicit social cognition research, which does not routinely address questions of morality. Moreover, the present findings add to a growing body of literature suggesting that implicit evaluations can flexibly encode the output of high-level inferential reasoning traditionally assumed to uniquely characterize controlled thought.

Open practices

All raw data files, analysis scripts, and materials used in this article are available for download from the Open Science Framework (OSF; https://osf.io/nt596/?view_only=e69c430c329449c1b897edc7232447fa). Study 2 was preregistered (https://aspredicted.org/blind.php?x=bb4ns6).

References

Bar-Anan, Y., & Nosek, B. A. (2017, March 5). A comparison of the sensitivity of four indirect evaluation measures to evaluative information. https://doi.org/10.31235/osf.io/edw6z/.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*(8), 1264–1272. https://doi.org/10.1037/xge0000383.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. WyerJr., & T. K. Srull (Eds.). *Handbook of social cognition: Basic processes; applications* (pp. 1–40). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bosson, J. K., Vandello, J. A., & Buckner, C. E. (2018). *The psychology of sex and gender.* Los Angeles, CA: SAGE Publications.

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition, 126*(2), 149–155. https://doi.org/10.1016/j.cognition.2012.09.006.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350. https://doi.org/10.1177/1088868312440047.

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2015). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience, 11*(3), 476–484. https://doi.org/10.1093/scan/nsv131.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014.

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences, 116*(20), 9802–9807. https://doi.org/10.1073/pnas.1903222116.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in experimental social psychology* (pp. 131–199). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2017.03.001.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8*(7), 342–353. https://doi.org/10.1111/spc3.12111.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology,* 127–183. https://doi.org/10.1016/bs.aesp.2019.09.004 Elsevier Inc.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037//0022-3514.56.1.5.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*(4), 574–595. https://doi.org/10.1177/1745691619826015.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in experimental social psychology* (pp. 59–127). Elsevier Inc. https://doi.org/10.1016/B978-0-12-385522-0.00002-0.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037//0033-295X.102.1.4.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216. https://doi.org/10.1037//0022-3514.85.2.197.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037//0033-295X.108.4.814.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science, 17*(10), 847–853. https://doi.org/10.1111/j.1467-9280.2006.01793.x.

Hester, N., & Gray, K. (2020). The moral psychology of raceless, genderless strangers. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691619885840 Advance online publication.

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin, 45*(2), 196–208. https://doi.org/10.1177/0146167218781340.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences, 9*(8), 357–359. https://doi.org/10.1016/j.tics.2005.06.011.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675–709. https://doi.org/10.1037/pspa0000046.

Kurdi, B., & Dunham, Y. (2019). *Sensitivity of implicit cognition to accurate and erroneous propositional inferences.* (Manuscript submitted for publication).

Kurdi, B., Morris, A., & Cushman, F. A. (2020, February 1). The role of causal structure in implicit cognition. https://doi.org/10.31234/osf.io/r7cfa.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., et al. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569–586. https://doi.org/10.1037/amp0000364.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823–849. https://doi.org/10.1037/pspa0000021.

Martin, J., Buon, M., & Cushman, F. A. (2019, August 2). The effect of cognitive load on intent-based moral judgment. https://doi.org/10.31234/osf.io/em9gx.

McLoughlin, N., & Over, H. (2017). Young children are more likely to spontaneously attribute mental states to members of their own group. *Psychological Science, 28*(10), 1503–1509. https://doi.org/10.1177/0956797617710724.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*(2), 183–198. https://doi.org/10.1017/S0140525X09000855.

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion, 27*(4), 743–752. https://doi.org/10.1080/02699931.2012.732040.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60*(C), 157–162. https://doi.org/10.1016/j.jesp.2015.05.009.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34*(5), 435–461. https://doi.org/10.1521/soco.2016.34.5.435.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37*(4), 557–569. https://doi.org/10.1177/0146167211400423.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.). *Classical conditioning II: Current research and theory* (pp. 64–99). (New York, NY).

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2006). Implicit and explicit attitudes respond differently to increasing amounts of counter-attitudinal information. *European Journal of Social Psychology, 37*(5), 867–878. https://doi.org/10.1002/ejsp.393.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1.

Sue, D. W. (2004). Whiteness and ethnocentric monoculturalism: Making the "invisible" visible. *American Psychologist, 59*(8), 761–769. https://doi.org/10.1037/0003-066X.59.8.761.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005.

van der Toorn, J., Ellemers, N., & Doosje, B. (2015). The threat of moral transgression: The impact of group membership and moral opportunity. *European Journal of Social Psychology, 45*(5), 609–622. https://doi.org/10.1002/ejsp.2119.

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology, 63*(C), 1–9. https://doi.org/10.1016/j.jesp.2015.11.002.

Van Dessel, P., Gawronski, B., & De Houwer, J. (2019). Does explaining social behavior require multiple memory systems? *Trends in Cognitive Sciences, 23*(5), 368–369. https://doi.org/10.1016/j.tics.2019.02.001.

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology, 69*(C), 23–32. https://doi.org/10.1016/j.jesp.2016.10.004.

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science, 10*(2), 266–273. https://doi.org/10.1177/1948550617752064.

Young, L., Campodron, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences, 107*(15), 6753–6758. https://doi.org/10.1073/pnas.0914826107.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005.

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology, 67*(11), 2105–2122. https://doi.org/10.1080/17470218.2014.907324.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67*(5), https://doi.org/10.2307/1131635 2478–16.