Cognition 225 (2022) 105116

Contents lists available at ScienceDirect

Cognition

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

The role of causal structure in implicit evaluation

Benedek Kurdi [a,*,1], Adam Morris [b], Fiery A. Cushman [b]

[a] Department of Psychology, Yale University, New Haven, CT, United States of America
[b] Department of Psychology, Harvard University, Cambridge, MA, United States of America

ARTICLE INFO

Keywords:
Associative learning
Causal learning
Dual-process theories
Implicit cognition
Implicit evaluation

ABSTRACT

Causal relationships, unlike mere co-occurrence, allow humans to obtain rewards and avoid punishments by intervening on their environment. Accordingly, explicit (controlled) evaluations of stimuli encountered in the environment are known to be sensitive to causal relationships above and beyond mere co-occurrence. In this project, we conduct stringent tests of whether implicit (automatic) evaluation also reflects causal relationships and begin to probe the representational mechanisms underlying such sensitivity. Participants (N = 4836) observed causal events during which two stimuli were equally contingent with positive or negative outcomes but only one of them was causally responsible for these outcomes. Across 6 studies, varying in design and amount of verbal scaffolding provided, differences in causal status consistently guided not only explicit measures of evaluation (Likert and slider scales; Bayes Factor meta-analysis: Cohen's $d = 0.28$, $BF_{10} > 10^{46}$) but also their implicit counterparts (Implicit Association Tests; Bayes Factor meta-analysis: Cohen's $d = 0.22$, $BF_{10} > 10^{29}$). However, unlike their explicit counterparts, implicit evaluations were not sensitive to causal relationships that had to be flexibly derived by combining disparate past experiences. Taken together, these studies suggest that implicit evaluations are sensitive to causal information. Such sensitivity seems to be mediated via precompiled, causally informed value representations rather than online computations over a causal model.

1. Introduction

Parents of newborns tend to keep diapers around, and they also tend to be less happy than their peers (Blanchflower, 2009). So, if you and your partner are hoping for a successful conception, here are some tips: Buy some diapers, and try not to have too much fun.

We all know better, and for a good reason: Humans represent not just statistical relationships between variables in the world, but also the causal relations among them (Buckner, Andrews-Hanna, & Schacter, 2008; Doll, Duncan, Simon, Shohamy, & Daw, 2015; Doll, Simon, & Daw, 2012; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; as do nonhuman animals, Tolman, 1949). The logic of cause and effect enables us to explain the past, predict the future, plan our actions, and assign responsibility and blame. It tells us that having babies causes parents to buy diapers, but buying diapers doesn't cause parents to have babies.

In the present project we examine whether automatic cognition, and more specifically implicit evaluation, is in on the joke. Specifically, we use implicit evaluations as a test case to ask whether automatic cognition (Kahneman, 2011; Sloman, 1996) is sensitive to representations of causal structure.[2] Automatic cognition is routinely indexed via implicit measures (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995; Meyer & Schvaneveldt, 1971) and other forms of suboptimal processing, such as cognitive load or time pressure (e.g., Phillips & Cushman, 2017; Rand, 2016). Here we rely on a widely validated

* Corresponding author at: Department of Psychology, Yale University, New Haven, CT 06511, United States of America.
E-mail address: benedek.kurdi@yale.edu (B. Kurdi).

[1] Author Note: Parts of this research have been presented at the 41st Annual Meeting of the Cognitive Science Society, in Montreal, Canada, in July 2019, and at the Annual Meeting of the Person Memory Interest Group, Severn Bridge, Canada, in October 2019. 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/afwyk/). Studies 1B, 5, and 6 were formally preregistered. We thank Elad Zlotnick and Victor Yang for assistance with study coding.

[2] Implicit evaluations and other aspects of automatic cognition may or may not behave the same in this regard. Specifically, the present results may not generalize to implicit memory or implicit semantic associations. In addition, we note that the present studies cannot speak to mechanisms of encoding: Given that we did not measure or manipulate the automaticity of the learning process (including aspects such as cognitive load, time pressure, or intentionality), the question that we can and do ask is whether causal knowledge, however it was encoded, is reflected by implicit (automatically retrieved) evaluations and, if so, what types of representations mediate this effect.

https://doi.org/10.1016/j.cognition.2022.105116
Received 19 February 2021; Received in revised form 24 March 2022; Accepted 28 March 2022
Available online 6 April 2022
0010-0277/© 2022 Elsevier B.V. All rights reserved.

measure of implicit evaluation, the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), to address the question of sensitivity to causal structure.

Current theories in psychology (De Houwer, 2014, 2018; Kurdi & Dunham, 2020; Lieberman, Gaunt, Gilbert, & Trope, 2002; Sloman, 1996, 2014; Smith & DeCoster, 2000; Strack & Deutsch, 2004) and philosophy (Gendler, 2008; Madva, 2016; Mandelbaum, 2016) diverge on this fundamental question. Moreover, existing empirical evidence is mostly indirect and ambiguous (Bading, Stahl, & Rothermund, 2020; De Houwer & Vandorpe, 2010; Gawronski & Brannon, 2021; Heycke & Gawronski, 2020; Hu, Gawronski, & Balas, 2017; Hughes, Ye, Van Dessel, & De Houwer, 2018; Moran & Bar-Anan, 2013; Moran, Bar-Anan, & Nosek, 2015).

Against this background of mixed existing evidence and opposing theoretical positions, the present experiments demonstrate that participants' causal representations influence their automatic judgments. In addition, the present experiments also constrain theories of how causal information reflected by implicit evaluations is represented in memory. Some current theories propose that automatic processes perform computations organized around causal principles and do so over explicit representations of causal relations (De Houwer, 2018; Mandelbaum, 2016; Sloman, 2014). An alternative theory is that such computations are not themselves a part of implicit evaluation, but instead generate precompiled representations that can be quickly retrieved during implicit evaluation (Kurdi & Dunham, 2020). In support of the second possibility, we find that implicit evaluations (unlike explicit ones) do not spontaneously integrate two elements of causal knowledge presented at separate time points (A → B, B → C) to derive a novel causal relation at evaluation time (A → C). This finding suggests that the processes subserving implicit evaluation do not operate over a causal model, but instead rely on causally informed representations compiled during prior learning.

1.1. Against the role of causal structure in implicit evaluation

Human cognition, including evaluation, involves many different kinds of process. Some are relatively fast, automatic, and unconscious, whereas others are relatively slow, effortful, and conscious (Greenwald & Banaji, 1995; Lieberman et al., 2002; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004). The first type of process is routinely indexed via implicit measures involving time pressure, cognitive load, or misattribution, whereas the second type of process is routinely indexed via self-report. There is considerable interest in the organizational principles underlying this division. A recurrent theme, spanning decades of theorizing, is that automatic cognition, and more specifically implicit evaluation, emerges from associative representations, while controlled cognition, and more specifically explicit evaluation, can draw on causal representations.

In his seminal review, Sloman (1996) proposes that "[r]ather than trying to reason on the basis of an underlying causal or mechanical structure, [implicit evaluation] constructs estimates based on underlying statistical structure" (p. 4). Similarly, according to Lieberman et al. (2002), when implicit evaluation is confronted with a causal inference problem, it "[…] simply estimates the similarity or associative strength between the antecedent and the consequent" (p. 210) without taking into account the direction of the causal relationship or the presence of confounding variables (see also Smith & DeCoster, 2000; Strack & Deutsch, 2004). Complementary arguments have also been made in philosophy, perhaps most prominently by Gendler (2008) who posits a distinction between purely associative and arational aliefs (implicit evaluation) and propositional and rational beliefs (explicit evaluation). Similarly, Madva (2016) contends that implicit evaluations are uniquely sensitive to "spatiotemporal relations in thought and perception" (p. 2659) but do not reflect how stimuli are related to each other, including their participation in causal relationships.

The claim that implicit evaluation is subserved by associative representations, and is insensitive to causal relationships, has some direct experimental support (Moran et al., 2015; Moran & Bar-Anan, 2013). In these studies, one family of novel creatures were repeatedly shown to cause a pleasant auditory stimulus to start, while a second family of creatures were repeatedly shown to cause that stimulus to stop; a third and a fourth family of creatures started or stopped unpleasant stimuli. If a cognitive process is sensitive only to statistical information and not to causal relationships, both creature families co-occurring with the pleasant stimulus should be evaluated positively and both creature families co-occurring with the unpleasant stimulus should be evaluated negatively. By contrast, if a cognitive process also reflects causality, then the creatures ending, rather than starting, stimuli should be evaluated opposite of the valence of the stimulus with which they had been paired.

Moran and Bar-Anan (2013) obtained the former (associative) pattern of results on implicit (indirect) measures, including the Implicit Association Test (IAT; Greenwald et al., 1998) and the Sorting Paired Features Task (SPF; Bar-Anan, Nosek, & Vianello, 2009) and the latter (causal) pattern of results on explicit (self-report) measures. However, in a subsequent experiment conducted by the same authors, implicit evaluations did show sensitivity to causal relations; this occurred when participants were instructed to form impressions of the four creatures rather than to memorize statistical relationships (Moran et al., 2015).

Moreover, in a different set of follow-up studies conducted by Bading et al. (2020), whether implicit evaluations reflected causal information depended on the comparisons investigated: When, similar to Moran and Bar-Anan (2013), the IAT contrasted two families that shared the same co-occurring stimulus valence (positive vs. negative) but differed in the type of action performed (start vs. stop), implicit evaluations seemed insensitive to causal status. However, when two families were contrasted that shared type of action performed (start vs. stop) but differed in co-occurring stimulus valence (positive vs. negative), then participants' IAT performance showed effects of causal information.

Similarly, Hu et al. (2017) obtained mixed evidence on the sensitivity of implicit evaluations to causal information in an evaluative conditioning procedure in which participants were exposed to pairings of drug brand conditioned stimuli (CSs) and positive and negative symptom unconditioned stimuli (USs). In this paradigm, verbal information on causal relations (e.g., the CS causing vs. preventing the US) influenced evaluative conditioning effects measured by the evaluative priming task (EPT; Fazio, Jackson, Dunton, & Williams, 1995) only when causal information was presented directly in conjunction with the relevant stimulus pairings. When causal information preceded the entire evaluative conditioning procedure, it influenced only explicit, but not implicit, evaluation of the CSs.

A final body of work also casts doubt on the role of causal information in implicit evaluation (e.g., Le Pelley, Griffiths, & Beesley, 2017; Rehder, 2009, 2014, 2015; Rehder & Waldmann, 2017). Specifically, this work demonstrates that purely associative, non-causal representations can interfere with causal reasoning. Notably, such interference can occur even when the measure used to probe causal knowledge is explicit, such as a self-report measure that gives participants ample time to carefully deliberate about their response. For example, in Rehder (2014), participants were asked to make judgments about the status of an outcome O after receiving information about two other variables (A and B) under three different causal models: common cause (A causing both B and O), chain inference (A causing B causing O), and common effect (A and B each causing O). Although under normative expectations involving the so-called causal Markov condition, participants should have reasoned differently about the outcome in each of the three cases, close to one third of participants reasoned purely associatively. That is, they showed equivalent patterns of responding in the common cause, chain inference, and common effect conditions. Moreover, even those participants who were generally sensitive to the causal status of the variables exhibited at least occasional interference via associative, non-causal processes. These findings would be naturally explained if even explicit causal reasoning was influenced by an associative, non-causal

implicit mechanism (e.g., Le Pelley et al., 2017).

1.2. In favor of the role of causal structure in implicit evaluation

In contrast to the prevailing view that automatic cognition operates exclusively over associative representations, several more recent theories propose that causal representations can also influence automatic cognitive processes, including implicit evaluation. Two decades after his seminal review of automaticity and control in human cognition, Sloman (2014) suggested that a revision was in order: "[t]he fact that we can recognize […] causal patterns quickly and automatically suggests that this information is encoded in [implicit evaluation]" (p. 76). Similarly, De Houwer (2018) in psychology and Mandelbaum (2016) in philosophy have taken the position that explicit and implicit evaluation are subserved by a common set of propositional representations that are sensitive to causal relations (among others).

Some initial evidence for this view was provided by De Houwer and Vandorpe (2010). In this study, participants were trained either that (*a*) two compounds produced identical health outcomes in isolation (A and B) and as a compound (AB), or that (*b*) two compounds produced different health outcomes in isolation (A and B) vs. as a compound (AB). This learning was applied to novel cases at test. Specifically, shown that a new compound (EF) produced a particular health outcome, both implicit and explicit judgments generalized from training the appropriate expectation: either that the two isolates (E and F) would produce the same outcome, or that they would produce a different one. Although this pattern of results is certainly consistent with an underlying causal representation, it could also be supported by appropriately structured statistical (non-causal) representations. Moreover, given that this experiment measured semantic and not valenced associations, its results may not generalize to implicit evaluation.

Further evidence for the potential sensitivity of implicit evaluations to causal structure was provided by Heycke and Gawronski (2020) and Gawronski and Brannon (2021). These studies implemented the same modified evaluative conditioning paradigm as the one used by Hu et al. (2017) and described above in which verbal labels (such as "causes" vs. "prevents") were used to modify the meaning of CS–US pairings. However, unlike in Hu et al. (2017), in these two papers the effects of causal information were measured using speeded self-report as the implicit measure. Corroborating the findings of Hu et al. (2017), implicit evaluations were found to reflect not only the valence of co-occurring stimuli but, in addition, also their causal status (cause vs. prevent).

Against this background, the present work makes several novel contributions. First, most prior work (including work by Hughes et al., 2018, discussed in more detail below) presented participants with causal information purely verbally, which several theoretical accounts suggest might favor propositional formats of representation. The present studies instead present information predominantly via direct observation rather than purely verbally, which is more likely to produce purely associative effects on implicit evaluation. Second, prior work contrasted causal and preventative relations with each other (e.g., A *leads to* B vs. A *prevents* B). An effect of this type of information on implicit evaluations could be explained by modified versions of associative accounts that allow for "validity tags" to reverse the meaning of valenced associations (e.g., Petty, Briñol, & DeMarree, 2007; Petty, Tormala, Briñol, & Jarvis, 2006). The present work instead contrasts causal relations with mere associations or stimuli participating in the same causal relationship either as a cause or as an effect. As such, in the present studies, the effects of causal information on implicit evaluation are more difficult to explain using enhanced versions of associative accounts.

Finally, the perhaps most relevant previous test of the role of causal information in implicit evaluation, which served as the starting point of the present work, was conducted by Hughes et al. (2018). In this study, participants viewed two faces, one followed by positive adjectives and the other followed by negative adjectives. Participants were told one of three things about the faces: that they were the faces of people who chose to show the adjectives to them (i.e., who "caused" the adjectives); that they were faces that the experimenter randomly chose as a predictive cue; or that they were unrelated to the adjectives. The difference in implicit evaluations between the two targets was largest in the causal condition, followed by the predictive and unrelated conditions, which produced equivalent effects. The difference across these two sets of conditions suggests that implicit evaluation may be sensitive to causal structure beyond mere statistical relatedness.

How would these results generalize beyond the specific case of an intentionally malicious agent? Potentially, the manipulation of the agent's causal responsibility for the negative adjectives affected implicit evaluations not directly, but instead through a representation of the agent's moral status, personality, or "value" more broadly. To illustrate one of these possibilities more concretely, the human actors introduced in this study probably also gave rise to inferences about hidden mental states such as beliefs and intentions ("Why did this guy choose such nasty adjectives for me? Presumably because he wanted me to feel bad."). And, as demonstrated by Kurdi, Krosch, and Ferguson (2020), mental state inferences about negative intentions are sufficient to produce negative implicit evaluations of human actors even in the absence of any adverse effects in the external world. As such, in the present experiments we instead presented participants with sparse, canonical causal relationships in a nonsocial setting.

In addition, the results by Hughes et al. (2018) are ambiguous in two further regards. First, the stronger difference between positive and negative targets in the causal condition relative to the predictive and unrelated conditions may have been due to causal relationships attracting closer attention or deeper processing than noncausal relationships in general, rather than specifically due to the causal role played by the human target. In the present project, we have the opportunity to address and provide evidence against this possibility. Second, as explained in more detail in the next section, the study by Hughes et al. (2018) does not address whether implicit evaluation can compute causal relations on the fly, or it merely accesses information precompiled on the basis of causal knowledge. In the current work, we also take up this issue.

1.3. Two potential cognitive mechanisms

To preview the results that we obtained, the first several studies presented here show that implicit evaluations are sensitive to causal relationships—a result difficult to reconcile with dual-process accounts that posit the existence of a purely associative automatic system in human thought (Gendler, 2008; Lieberman et al., 2002; Madva, 2016; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004). While several current theories can accommodate this finding, they disagree on the timing and location of the relevant causal computations.

One possibility is that computations over explicitly causal principles occur within the processes responsible for implicit evaluation. For instance, Sloman (2014) holds that automatic cognition, including implicit evaluation, is capable of consulting and performing novel operations over a causal model, and De Houwer (2018) and Mandelbaum (2016) claim that implicit evaluation operates over propositional representations. Although these two proposals differ in subtle details, from the perspective of the current work they converge on the idea that implicit evaluations can be updated flexibly as causal information about the environment emerges. Below we refer to this perspective as the "causal computation" account to express the idea that automatic processes can implement causal principles to compute new causal relations.

Alternatively, it may be that the processes responsible for implicit evaluation do not compute over causal principles directly, but instead have access to representations that are precompiled in a causally informed way during learning or offline thinking, and which are then automatically retrieved at test. For instance, a competitive skier who repeatedly visualizes her slalom run may deliberatively apply causal principles during visualization to precompile a set of automatic

reactions that she will deploy during the race. The automatic processes that implement these precompiled representations would then be sensitive to causal content without computing it directly. Some evidence implicates processes of this kind in human value-guided decision-making (Gershman, Markman, & Otto, 2014) and in causal reasoning (Fernbach & Sloman, 2009).

In a similar vein, Kurdi and Dunham (2020) recently proposed that implicit and explicit evaluation are sensitive to the same sources of information, including reasoning about causes and effects. However, they may rely on representations that capture such information in fundamentally different ways. Specifically, explicit evaluation is thought to draw upon all the details of the full causal model (e.g., A causes B and B causes a negative outcome O_1; A is correlated with C; C does not cause B). This architecture enables flexible, productive causal reasoning. By contrast, implicit evaluation is thought to access only the ultimate evaluative implications of this causal model [value(C) > value(A)]. In other words, implicit evaluation has access to the products of causal computations; however, once created, those products do not contain explicit symbols for that causal knowledge. Below we refer to this view as the "causal access" account.

In the final study of the present project, we used a revaluation paradigm (Adams & Dickinson, 1981; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gershman et al., 2014; Kurdi, Gershman, & Banaji, 2019) to newly contrast the predictions of these two theoretical positions. In this paradigm, participants first form a value representation of a stimulus based on its causal affordances, but later receive evidence that the affordances supporting that value representation no longer obtain. If implicit evaluation operates directly over a causal model, then participants should be able to flexibly integrate such new evidence into the representation underlying their implicit evaluation of the causal stimulus (just as they do for their explicit evaluation). Conversely, if implicit evaluation is subserved by cached representations (and does not compute over a causal model online), then such an intervention may leave preexisting implicit evaluations in place.

1.4. The present project

In the first five experiments reported in this paper we tested whether implicit evaluation reflects causal structure ("causal hypothesis") or is sensitive only to co-occurrence information. ("associative hypothesis"). These causal events involved inanimate physical objects and mechanical devices, allowing us to carefully equate actual and perceived co-occurrences between stimuli and minimize unnecessary differences between conditions.

Specifically, in Studies 1–4, participants observed the operation of a causal system in which some objects caused positive or negative outcomes, while other objects merely co-occurred with, but were not causally responsible for, these outcomes. We started out by testing the basic effect in the presence (Studies 1A and 1B) and absence (Study 2) of detailed verbal instructions about the operation of the causal system. In Study 3, we replicated the effect while eliminating temporal primacy and attentional capture by valenced stimuli as alternative explanations. In Study 4 we investigated sensitivity of the effect to the valence of the outcome. Finally, in Study 5, initially neutral stimuli participated in a causal relationship in both between-subject conditions but in one as a cause and in the other as the effect, thus providing an especially stringent test of the role of causal status.

In Study 6, we investigated the cognitive mechanism underlying the effect. Specifically, we examined whether differences in implicit evaluation arising from differences in causal status are amenable to updating as a result of revaluing an initially negative outcome. Under recent proposals by De Houwer (2018), Mandelbaum (2016), and Sloman (2014), implicit evaluation is subserved by propositional representations or online computations over a causal model; as such, implicit evaluations should be flexibly updated as a result of changes to the causal model (the "causal computation" view). By contrast, if, as suggested by Kurdi and Dunham (2020), the outputs of causal reasoning are precompiled into a compressed value representation during learning, implicit evaluations should be sensitive to initial differences in causal status but should not reflect learning that requires access to details of the already represented causal model (the "causal access" view).

2. Studies 1A and 1B

In these studies, we conduct an initial test of whether implicit (and explicit) evaluations reflect causal structure above and beyond the effects of co-occurrence information. Specifically, participants were exposed to a series of physical events in the course of which one set of stimuli was causally responsible for a positive outcome, whereas a different set of stimuli was equally predictive of, but not causally responsible for, that outcome. (In fact, co-occurring stimuli appeared closer to the positive stimulus than causal stimuli in both space and time.) All relevant theories predict that explicit evaluations should be sensitive to the difference in causal status. However, predictions differ with regard to implicit evaluations: Under the associative hypothesis, implicit evaluations are posited to track only statistical relatedness but not causal structure; under the causal hypothesis, both explicit and implicit evaluation are predicted to reflect differences in causal status.

2.1. Method

2.1.1. Participants

Participants in both studies were adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu). We report these two studies together because their procedures were nearly identical, with one small change described below. For Study 1B, the sample size and exclusion criteria, analytic choices, and predictions were preregistered (https://aspredicted.org/5yr69.pdf).

404 participants were recruited in Study 1A and 518 in Study 1B. Participants who did not complete the Implicit Association Test (IAT; Greenwald et al., 1998), which served as the main dependent measure ($n = 7$ in Study 1A and $n = 6$ in Study 1 B), and participants whose response latencies were below 300 ms on at least 10% of IAT trials ($n = 4$ in Study 1A and $n = 3$ in Study 1B) were excluded from further analyses (Greenwald, Nosek, & Banaji, 2003). Thus, participant exclusions resulted in final sample sizes of 393 and 509, respectively.

The sample in Study 1A consisted of 263 women and 127 men; one participant did not report their gender. The mean age in Study 1A was 41 years ($SD = 15$ years). The sample in Study 1B consisted of 361 women and 147 men; one participant did not report their gender. The mean age in Study 1B was 42 years ($SD = 16$ years). Further demographic variables, including race, education, and political ideology, are available for secondary analyses in the open data for this and all remaining studies.

2.1.2. Materials

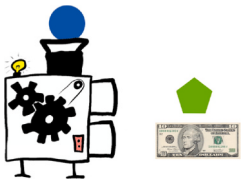
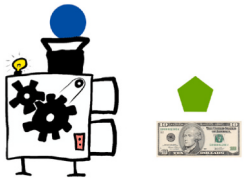
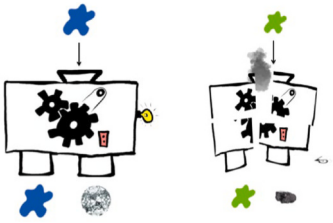
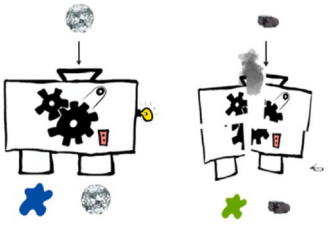
An image of a machine with an activator surface and a light bulb on top and two emission slots attached to its right side served as the setting for causal events to occur (see Table 1). An image of a $10 bill served as the positive stimulus. Five shapes (a square, a circle, a triangle, a pentagon, and a blob) of three colors each (blue, green, and purple) served as the causal, co-occurring, and inert stimuli, respectively (see below).

2.1.3. Procedure and measures

The study consisted of a learning phase and a test phase. In the learning phase, participants learned about the operation of a machine, including the difference between causal and co-occurring stimuli. In the test phase, they completed an Implicit Association Test (IAT; Greenwald et al., 1998), followed by explicit measures of memory and evaluation.

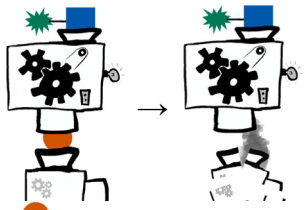
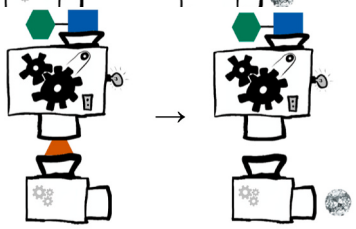
At the outset of the learning phase, participants were instructed that

Table 1
Overview of the paradigm of Studies 1–6. For the purposes of demonstration, blue shapes serve as causal stimuli and green shapes always serve as co-occurring stimuli in Studies 1–4 and 6, and blue shapes serve as positive stimuli and green shapes serve as negative stimuli in Study 5. In the studies, the assignment of colors to causal roles was counterbalanced. "Main trial(s) of interest" refers to the trial or trials involving the stimuli that were subsequently tested on the IAT. In Study 6, adjacent displays show the temporal progression of each trial type, which is marked by an arrow (→).

Study	Main trial(s) of interest	Number and type of trials	Note
Studies 1A and 1B		10 causal trials (shown) 10 inert trials involving third color	Trials presented in randomized order
Study 2		10 causal trials (shown) 10 inert trials involving third color	Trials presented in randomized order
Study 3		12 positive causal trials (shown) 4 inert trials involving the same colors but opposite orientation 4 negative causal trials involving third and fourth color	The first two and last two trials were always causal, with the order of the remaining trials randomized
Study 4 Positive condition		10 positive causal trials (shown) 3 inert trials involving the same colors but opposite orientation 3 negative causal trials involving third and fourth color	The first and last two trials were always positive causal, with the order of the remaining trials randomized
Study 4 Negative condition		10 negative causal trials (shown) 3 inert trials involving the same colors but different orientation 3 positive causal trials involving third and fourth color	The first trial was always positive causal and last two trials were always negative causal, with the order of the remaining trials randomized
Study 5 Cause condition		8 positive causal trials (shown left) 8 negative causal trials involving second color (shown right)	The order of trials was fully randomized; colored shapes were shown to enter the machine through the slot above
Study 5 Effect condition		8 positive effect trials (shown left) 8 negative effect trials involving second color (shown right)	The order of trials was fully randomized; valenced stimuli were shown to enter the machine through the slot above

(*continued on next page*)

Table 1 (*continued*)

Study	Main trial(s) of interest	Number and type of trials	Note
Study 6 Baseline, revaluation, and relearning conditions (part 1)		8 negative causal trials (shown) 4 inert trials involving the same colors but different orientation	The first two trials were always causal trials, with the order of the remaining trials randomized
Study 6 Revaluation condition (part 2)		12 positive trials not involving the original causal and co-occurring colors (shown)	
Study 6 Relearning condition (part 2)		8 positive causal trials involving the original causal and co-occurring colors (shown) 4 inert trials involving the same colors but different orientation	The first two trials were always causal trials, with the order of the remaining trials randomized

they would learn about an automatic dispenser. They were informed that the dispenser would dispense two things, objects of a certain color (co-occurring stimuli) and money (the positive stimulus). However, to be activated, objects of a different color (causal stimuli) would have to be put on top of the machine, whereas activators of any other color (inert stimuli) would not activate the machine. To avoid unduly creating a contingency between causal stimuli and the positive stimulus in the verbal instructions, and thus biasing results in favor of the causal hypothesis, the two were never mentioned together in the same sentence, whereas the co-occurring stimuli and the positive stimulus were mentioned together twice. As such, any explicit or implicit preference in favor of the causal stimuli over the co-occurring stimuli cannot be attributed to a difference in verbal associations.

Following the initial instructions, participants were exposed to a total of 20 trials involving the machine, as well as the causal, co-occurring, and inert stimuli described above. Each of the three colors (blue, green, and purple) was randomly assigned to play the role of causal, co-occurring, and inert stimuli, respectively. The 20 trials included 10 causal trials and 10 inert trials, which were presented in individually randomized order (see Table 1). On causal trials, a randomly selected causal stimulus was shown to drop onto the activator surface of the machine, upon which the light bulb on top of the machine switched on, indicating that the machine was operational, and the machine emitted a randomly selected co-occurring stimulus and a positive stimulus. On inert trials, a randomly selected inert stimulus was shown to drop onto the activator surface of the machine; however, unlike on causal trials, the machine was not activated. Participants progressed from trial to trial at their own pace using the space bar.

In the test phase, participants completed an Implicit Association Test (IAT; Greenwald et al., 1998) measuring implicit evaluations of causal stimuli relative to co-occurring stimuli. The IAT was followed by two explicit memory items probing participants' recollection for the details of the learning phase and one explicit evaluation item for the causal, co-occurring, and inert stimuli each.[3] Given that the present project focuses primarily on implicit processes, the IAT was always administered first to be able to obtain a relatively pure measure of implicit evaluations, uncontaminated by having completed explicit measures previously.

Implicit evaluations of causal stimuli relative to the co-occurring stimuli 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 block in which one set of targets (e.g., causal stimuli) share a response key with positive items and the other set of targets (e.g., co-occurring stimuli) share a response key with negative items, and a second block in which the assignment of targets to valences is reversed (e.g., co-occurring–positive vs. causal–negative).

In block 1 (category practice block; 20 trials), participants used two response keys (E and I) to sort the causal and co-occurring stimuli encountered in the learning phase. Color names (e.g., "blue" and "green") were used as category labels. In block 2 (attribute practice block; 20 trials), participants sorted positive words ("glory," "happy," "joy," "love," "lucky," "peace," and "sweet") and negative words ("agony," "bitter," "bomb," "devil," "grief," "hate," and "war"). The words "good" and "bad" were used as attribute labels. In block 3 (congruent combined block; 40 trials), participants in Study 1A used one response key to sort causal stimuli and positive words and the other response key to sort co-occurring stimuli and negative words. In block 4 (reversed category practice block; 20 trials), participants sorted the same causal and co-occurring stimuli used in blocks 1 and 3 but with the mapping of categories to response keys reversed. Finally, in block 5 (incongruent combined block; 40 trials), participants in Study 1A used one response key to sort co-occurring stimuli and positive words and the other response key to sort causal stimuli and negative words. In Study

[3] The explicit memory items were included for exploratory purposes and will not be discussed further for this or any of the remaining studies. The text of these items as well as results involving these items are available from OSF (https://osf.io/afwyk/).

1B, the order of the two combined blocks was counterbalanced given that presenting the congruent combined block first can artificially inflate the estimate of the difference in implicit evaluations.[4]

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 causal stimuli compared to the co-occurring stimuli, in line with the causal hypothesis.

Participants were also asked to indicate how positive or negative their feelings toward blue shapes, green shapes, and purple shapes were based on their experience in the study. For each item, the response options included "very positive" (coded as 5), "positive" (coded as 4), "neutral" (coded as 3), "negative" (coded as 2), and "very negative" (coded as 1). The order of the three items was individually randomized. The explicit evaluation score for co-occurring stimuli was subtracted from the explicit evaluation score for causal stimuli to create an index of relative preference and thus make explicit evaluations comparable to implicit evaluations.[5]

2.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 1. Descriptively, in line with the causal hypothesis, both measures seemed to exhibit a preference in favor of the causal stimulus over the co-occurring stimulus, as reflected by positive average explicit evaluations ($M = 0.31$, $SD = 1.12$ in Study 1A and $M = 0.30$, $SD = 1.00$ in Study 1B) and positive average implicit evaluations ($M = 0.21$, $SD = 0.52$ in Study 1A and $M = 0.12$, $SD = 0.56$ in Study 1B).

2.2.1. Explicit evaluations

Confirming the predictions derived from relevant theoretical accounts, explicit evaluations of causal stimuli were found to be significantly more positive than explicit evaluations of co-occurring stimuli, $t(384) = 5.41$, $p < .001$, $BF_{10} = 6.96 \times 10^4$, Cohen's $d = 0.28$, in Study 1A, and $t(505) = 6.79$, $p < .001$, $BF_{10} = 1.67 \times 10^8$, Cohen's $d = 0.30$, in Study 1B.

2.2.2. Implicit evaluations

In line with the causal hypothesis and paralleling the results obtained using explicit measures, implicit evaluations of causal stimuli were found to be significantly more positive than those of co-occurring stimuli, $t(390) = 7.94$, $p < .001$, $BF_{10} = 2.25 \times 10^{11}$, Cohen's $d = 0.40$, in Study 1A, and $t(508) = 4.67$, $p < .001$, $BF_{10} = 2.00 \times 10^3$, Cohen's $d = 0.21$, in Study 1B.

2.3. Discussion

Study 1 provides initial evidence for the causal hypothesis. Specifically, both explicit and implicit measures revealed more positive evaluations of stimuli causally responsible for a positive outcome than stimuli merely co-occurring with, but not causally responsible for, that outcome. To ascertain the robustness of this finding, Studies 2–5 offer additional, increasingly stringent, tests of the causal hypothesis.

3. Study 2

In the most pertinent previous studies of the effects of causal relationships on implicit evaluation (Gawronski & Brannon, 2021; Heycke & Gawronski, 2020; Hu et al., 2017; Hughes et al., 2018), causal information was communicated exclusively via verbal instructions. In Study 1, we provided participants with both verbal instructions and an extensive direct observation period illustrating the relevant causal relations. However, much causal learning unfolds without any explicit verbal instruction (e.g., Meltzoff, Waismeyer, & Gopnik, 2012). As such, the case for the sensitivity of implicit evaluation to causal status would be stronger if causally sensitive implicit evaluations could be formed through mere observation. To test this possibility, in Study 2, we replicated the basic design of Study 1 without any preceding verbal instructions about the operation of the causal system.

3.1. Method

3.1.1. Participants

The method of recruitment and exclusion criteria were identical to those used in Study 1. Of an initial sample of 355, six participants were excluded for failing to complete the IAT and three for excessively fast responding, resulting in a final sample size of 346. The sample consisted of 230 women and 114 men. The mean age was 36 years ($SD = 15$ years).

3.1.2. Procedure and measures

The procedure and measures were nearly identical to those implemented in Study 1, with three exceptions. First, unlike in Study 1, where participants received detailed verbal instructions about the operation of the machine, in Study 2 participants were merely informed that they would learn about an automatic dispenser and that they would see blue shapes, green shapes, purple shapes, and money over the course of the learning task. Second, unlike in Study 1A and similar to Study 1B, the order of the two critical IAT blocks was counterbalanced. Third, unlike in Study 1 where the three colors were randomly assigned to the roles of causal, co-occurring, and inert stimuli, in the present study green shapes always served as inert stimuli, and blue and purple shapes were randomly assigned to serve either as causal or as co-occurring stimuli.

3.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 2. Descriptively, similar to Study 1, both measures seemed to exhibit a preference in favor of the causal stimulus over the co-occurring stimulus, as reflected by positive average explicit evaluations ($M = 0.16$, $SD = 0.94$) and positive average implicit evaluations ($M = 0.07$, $SD = 0.49$).

3.2.1. Explicit evaluations

Similar to Study 1, explicit evaluations of causal stimuli were found to be significantly more positive than explicit evaluations of co-occurring stimuli, $t(339) = 3.07$, $p = .002$, $BF_{10} = 6.15$, Cohen's $d = 0.17$.

3.2.2. Implicit evaluations

Similar to Study 1 and in line with the causal hypothesis, implicit evaluations of causal stimuli were significantly more positive than those of co-occurring stimuli, $t(343) = 2.72$, $p = .007$, $BF_{10} = 2.27$, Cohen's $d = 0.15$. Given the more subtle nature of the manipulation, the difference in evaluation between causal stimuli and co-occurring stimuli was smaller than in Study 1 on both explicit and implicit measures.

3.3. Discussion

Study 2 provides further evidence in favor of the causal hypothesis. Specifically, we replicated the pattern of results obtained in Study 1: Stimuli shown to be causally responsible for a positive outcome were evaluated more positively on both explicit and implicit measures than stimuli equally co-occurring with but not causally responsible for that outcome. Notably, this result emerged in the absence of any prior verbal instructions containing relational information about differences in

[4] We thank Jan De Houwer for alerting us to this issue.

[5] For clarity of presentation, analyses involving explicit evaluations of stimuli not tested on the implicit measure are not reported in the main text for this or any of the remaining studies. However, these analyses can be downloaded as a supplement from OSF (https://osf.io/afwyk/). Across all studies, results on explicit measures were in line with theoretical predictions.

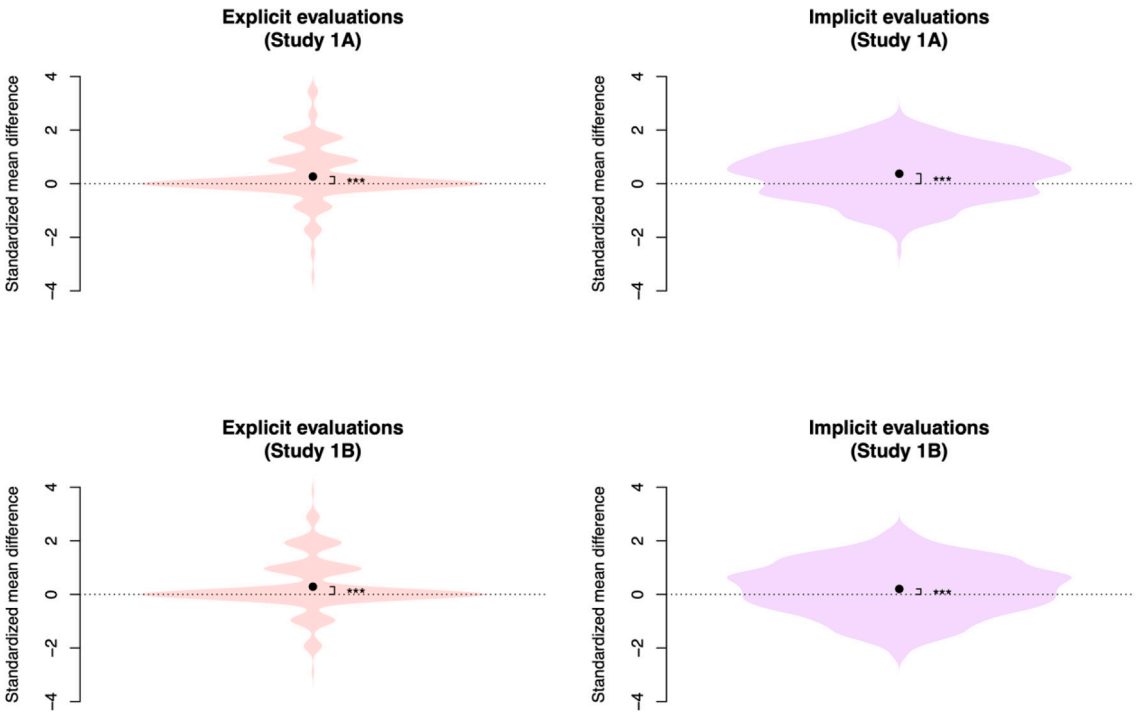


Fig. 1. Distribution of explicit and implicit evaluations (Study 1), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for the causal over the co-occurring stimulus. Differences relative to zero significant at $p < .001$ are marked ***.

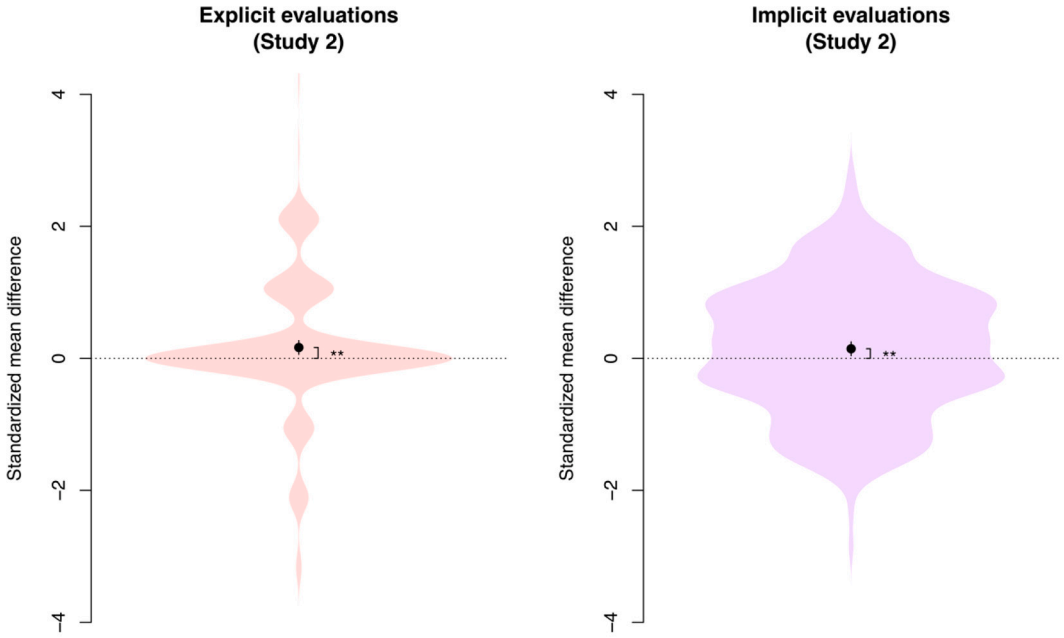


Fig. 2. Distribution of explicit and implicit evaluations (Study 2), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for the causal over the co-occurring stimulus. Differences relative to zero significant at $p < .01$ are marked **.

causal status. As such, the present study suggests that causal inferences can be made purely from observing the physical world, with such causal inferences subsequently influencing not only explicit but also implicit evaluation.

4. Study 3

Studies 1 and 2 jointly suggest that implicit evaluations can reflect causal relationships in addition to mere co-occurrence information, even when such causal relationships are inferred purely from observation. However, the particular causal events used in Studies 1 and 2 are subject to two alternative interpretations. Specifically, in Studies 1 and 2, causal stimuli always appeared on the screen before co-occurring stimuli did. As such, in line with the "first is best" effect (Carney & Banaji, 2012), implicit (and explicit) preferences in favor of the causal over the co-occurring stimulus may have emerged due to a difference in temporal

order of appearance rather than, as hypothesized, a difference in causal status. Moreover, in Studies 1 and 2, co-occurring stimuli always appeared on the screen simultaneously with the intrinsically positive stimulus. Given that positively valenced stimuli are known to bias visual attention (Theeuwes & Belopolsky, 2012), such attentional capture may have interfered with the encoding of the co-occurring stimulus for reasons unrelated to causal status.

To eliminate these potential confounds, in Study 3, causal and co-occurring stimuli were presented simultaneously, before the positive stimulus appeared on the screen. For instance, in one condition, participants watched a series of events over the course of which a blue and green object (connected to each other) were lowered onto a machine. However, only the blue object touched the machine's activator surface, thereby activating the machine and producing a positive outcome. Meanwhile, the green object remained inert. Thus, in this example, the blue and green objects shared nearly identical spatiotemporal relationships with positive stimuli, but only the blue object was causal.

4.1. Method

4.1.1. Participants

The method of recruitment and exclusion criteria were identical to those used in Studies 1 and 2. Of an initial sample of 355, five participants were excluded for failing to complete the IAT and seven for excessively fast responding, resulting in a final sample size of 343. The sample consisted of 232 women and 111 men. The mean age was 34 years ($SD = 15$ years).

4.1.2. Materials

The blue, green, and purple shapes used in Studies 1 and 2 were retained for use in Study 3. However, some changes were also implemented. First, an image of a different machine was used. This machine had the light bulb attached to its right-hand side rather than to its top and had one, rather than two, emission slots. The emission slot was placed at the bottom, rather than on the right-hand side, of the machine (see Table 1). Second, instead of the image of a $10 bill, the image of a diamond was used as an intrinsically positive stimulus. Third, an image of toxic sludge was newly added to be used as an intrinsically negative stimulus. The addition of this stimulus served as a pretest for Study 4 and is not discussed further. Fourth, in addition to shapes of the three colors used in Studies 1 and 2, five shapes of a fourth color (orange) were newly included.

4.1.3. Procedure and measures

The procedure and measures were similar to those implemented in Studies 1 and 2. Participants learned about the operation of a machine, including the difference between causal and co-occurring stimuli, in a learning phase, and completed an IAT, followed by explicit measures of memory and evaluation, in a test phase.

At the outset of the learning phase, participants were instructed that they would learn about a machine that is designed to make diamonds but, if not operated properly, makes toxic sludge. Then they received the following information:

> At the top right, the machine has an activator surface. Just the right kind of activator must be put on this surface to activate the machine. Specifically, the machine only works properly if a(n) [positive causal color] activator touches the activator surface. However, activators come in pairs. Sometimes the [co-occurring (inert) color] end of the activator touches the activator surface. In such cases, the machine doesn't do anything. Finally, if an activator of any other color touches the activator surface, the machine makes toxic sludge instead of a diamond.

Following the initial instructions, participants were exposed to a total of 20 trials involving the machine and the four sets of colored shapes described above. Shapes of one color were randomly assigned to serve as positive causal stimuli, and shapes of a second color were randomly assigned to serve as co-occurring (inert) stimuli, with the remaining two colors serving as negative causal stimuli. The 20 trials included 12 positive causal trials, four inert trials, and four negative causal trials. The first two trials were always positive causal trials, with the order of the remaining 18 trials individually randomized (see Table 1).

Importantly, positive causal stimuli and co-occurring (inert) stimuli were always shown to be tethered to each other. On the theoretically crucial positive causal trials, a randomly selected positive causal stimulus was shown to drop onto the activator surface of the machine, upon which the light bulb switched on, indicating that the machine was operational, and the machine emitted a positive stimulus. On these trials, a randomly selected co-occurring (inert) stimulus was shown to be connected to the positive causal stimulus but did not touch the activator surface of the machine. Inert trials involved the same stimuli as positive causal trials; however, instead of the positive causal stimulus, the co-occurring (inert) stimulus touched the activator surface of the machine and the machine was not activated. On negative causal trials, a randomly selected negative causal stimulus was shown to drop onto the activator surface of the machine, upon which the machine emitted a negative stimulus.

The test phase was identical to the test phase of Studies 1 and 2, with the sole exception that four, rather than three, explicit evaluation items were administered given the newly added fourth color.

4.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 3. Descriptively, similar to Studies 1 and 2, both measures seemed to exhibit a preference in favor of the positive causal stimulus over the co-occurring (inert) stimulus, as reflected by positive average explicit evaluations ($M = 0.81$, $SD = 1.21$) and positive average implicit evaluations ($M = 0.22$, $SD = 0.49$).

4.2.1. Explicit evaluations

In line with the predictions derived from relevant theoretical accounts and the results of Studies 1 and 2, explicit evaluations of positive causal stimuli were found to be significantly more positive than explicit evaluations of co-occurring (inert) stimuli, $t(340) = 12.48$, $p < .001$, $BF_{10} = 2.58 \times 10^{26}$, Cohen's $d = 0.68$.

4.2.2. Implicit evaluations

Similar to Studies 1 and 2 and in line with the causal hypothesis, implicit evaluations of causal stimuli were significantly more positive than those of co-occurring (inert) stimuli, $t(342) = 8.15$, $p < .001$, $BF_{10} = 7.31 \times 10^{11}$, Cohen's $d = 0.44$.

4.3. Discussion

Study 3 provides another demonstration of the sensitivity of implicit evaluations to causal status, with stimuli causally responsible for a positive outcome evaluated more positively than stimuli merely predictive of that outcome. Notably, under the temporal primacy and attentional capture accounts, the difference between causal and co-occurring stimuli should have been eliminated in the present study. Instead, the effect of causal status on implicit (and explicit) evaluation in

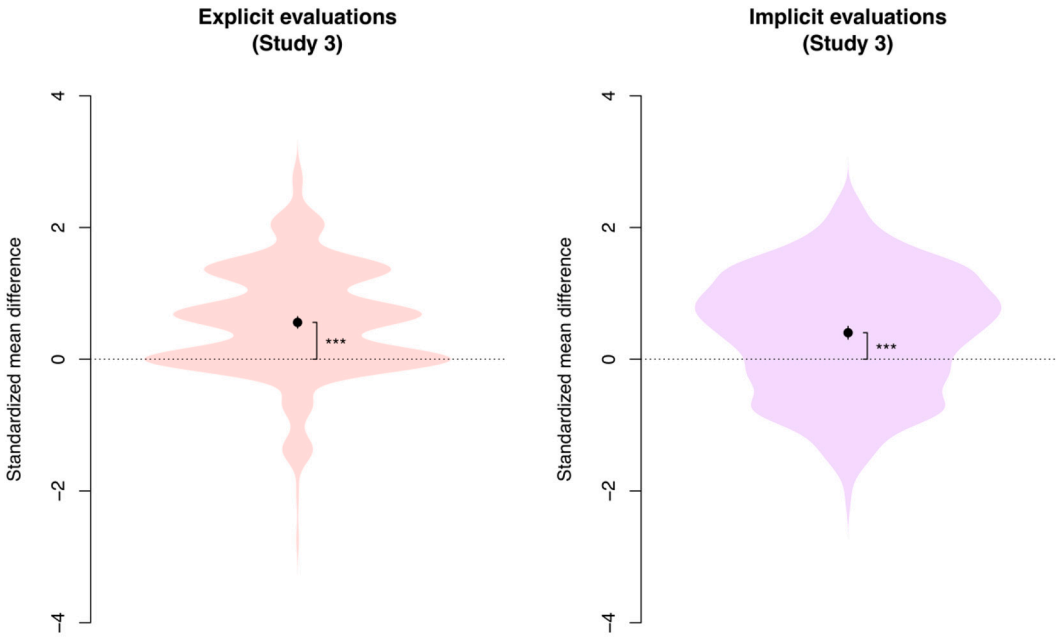


Fig. 3. Distribution of explicit and implicit evaluations (Study 3), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for the causal over the co-occurring (inert) stimulus. Differences relative to zero significant at $p < .001$ are marked ***.

Study 3 was descriptively stronger, rather than weaker, than in either previous experiment.[6]

5. Study 4

Studies 1–3 suggest that objects causing a positive outcome are evaluated more positively than objects merely co-occurring with that outcome. We interpret this as an indication that, in keeping with standard models of reinforcement learning, causes inherit the valence of their effects (Daw et al., 2011; Doll et al., 2012; Gershman et al., 2014). But because the implicit measures administered in Studies 1–3 involved only stimuli co-occurring with positive outcomes, it is conceivable that causally efficacious stimuli are always evaluated more positively than causally non-efficacious stimuli, irrespective of the valence of the outcome. If so, then an object that causes a negative outcome would be evaluated more positively than one co-occurring with that outcome because of a general evaluative "cause bonus."

Therefore, in Study 4, we tested whether outcome valence (positive vs. negative) modulates the effects of causal status on implicit evaluation. To do so, we modified the design of Study 3 so that in a newly added condition we focused on implicit evaluations of stimuli causally responsible for vs. merely co-occurring with the appearance of an undesirable, rather than desirable, object. We predicted that objects causing positive outcomes would be evaluated more positively than objects merely co-occurring with that outcome, and that objects causing negative outcomes would be evaluated more negatively than objects merely co-occurring with negative outcomes.

5.1. Method

5.1.1. Participants and design

The method of recruitment and exclusion criteria were identical to those used in Studies 1–3. Of an initial sample of 305, three participants were excluded for failing to complete the IAT and seven for excessively fast responding, resulting in a final sample size of 295. The sample consisted of 198 women and 97 men. The mean age was 37 years (SD = 14 years). Participants were randomly assigned to a positive condition (n = 168) or a negative condition (n = 127) in a between-subjects design.

5.1.2. Materials

The materials used in Study 3 were retained for use in Study 4. In addition, an image of the broken version of the machine as well as an image of smoke were created for use on negative causal trials (see Table 1).

5.1.3. Procedure and measures

The procedure and measures were similar to those implemented in Study 3. That is, participants learned about the operation of a machine, including the difference between causal and co-occurring stimuli, in a learning phase, and completed an IAT, followed by explicit measures of memory and evaluation, in a test phase.

At the outset of the learning phase, participants in both conditions were instructed that they would learn about a machine that is designed to make diamonds but, if not operated properly, breaks apart and emits toxic smoke. Similar to Study 3, participants in both conditions were informed that the causal stimuli were tethered to co-occurring (inert) stimuli. In both conditions, the inert stimuli were described as not causing any changes in the operation of the machine. Positive causal stimuli were described as causing the machine to emit diamonds and negative causal stimuli were described as causing the machine to break apart and to emit toxic smoke.

Following the initial instructions, participants were exposed to a total of 16 trials involving the machine and the four sets of colored shapes described above. The trial types were identical across the positive and negative conditions; however, their frequencies differed. Positive

[6] An anonymous reviewer suggested that the difference between the causal and associated stimulus might be larger in Study 3 than in previous studies because in this study, unlike in previous studies, the associated stimulus was shown to be causally inert. This stands out as a likely possibility, which depends on the claim that causal representations influence implicit evaluation above and beyond mere statistical associations.

causal trials were identical to the positive causal trials included in Study 3. Inert trials were also identical to the inert trials included in Study 3. In the positive condition, the inert trials showed causal positive stimuli and the co-occurring (inert) stimuli, whereas in the negative condition, the inert trials showed causal negative stimuli and co-occurring (inert) stimuli. In both conditions, the inert rather than the causal color made contact with the activator surface of the machine. Finally, negative causal trials were modified relative to Study 3. On these trials, a negative causal stimulus touched the activator surface of the machine, upon which the machine was shown to break into pieces and to emit smoke.

In the positive condition, participants were exposed to 10 positive causal trials involving positive causal stimuli and co-occurring (inert) stimuli, three inert trials involving the same stimuli but in the opposite orientation, and three negative causal trials involving stimuli of the third and fourth colors. The first and last two trials were always positive causal trials, with the order of the remaining trials individually randomized for each participant. In the negative condition, participants were exposed to 10 negative causal trials involving negative causal stimuli and co-occurring (inert) stimuli, three inert trials involving the same stimuli but in the opposite orientation, and three positive causal trials involving stimuli of the third and fourth colors. The first trial was always a positive causal trial and the last two trials were always negative causal trials, with the order of the remaining trials individually randomized for each participant.

The test phase was identical to the test phase of Studies 1–3, with the sole modification that the IAT measured implicit evaluations of positive causal stimuli relative to co-occurring (inert) stimuli in the positive condition and implicit evaluations of negative causal stimuli relative to co-occurring (inert) stimuli in the negative condition.

5.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 4. Descriptively, it seems that the valence of the outcome modulated both explicit and implicit evaluations. Specifically, explicit evaluations in the positive condition appeared to replicate results from Study 3 in that causal stimuli were preferred to co-occurring stimuli ($M = 0.79$, $SD = 1.16$). However, this preference was reversed in the negative condition where co-occurring stimuli were preferred to causal stimuli ($M = -0.50$, $SD = 1.31$). The same trend was apparent on the implicit measure, with a preference in favor of causal stimuli over co-occurring stimuli emerging in the positive condition ($M = 0.24$, $SD = 0.46$) and the opposite, but considerably weaker, preference manifesting itself in the negative condition ($M = -0.07$, $SD = 0.54$).

5.2.1. Explicit evaluations

We found a significant and large effect of condition (positive vs. negative), $t(252.48) = 8.85$, $p < .001$, $BF_{10} = 1.56 \times 10^{14}$, Cohen's $d = 1.06$, demonstrating the sensitivity of causally informed explicit evaluations to the valence of outcomes. We then investigated the difference in explicit evaluations between causal and noncausal stimuli separately in each condition. In the positive condition, explicit evaluations of causal stimuli were found to be significantly more positive than explicit evaluations of co-occurring (inert) stimuli, $t(167) = 8.87$, $p < .001$, $BF_{10} = 5.33 \times 10^{12}$, Cohen's $d = 0.68$, replicating the result obtained in Study 3. Conversely, in the negative condition, explicit evaluations of causal stimuli were found to be significantly more negative than explicit evaluations of co-occurring (inert) stimuli, $t(126) = -4.34$, $p < .001$, $BF_{10} = 535.49$, Cohen's $d = 0.39$.

5.2.2. Implicit evaluations

We found a significant and medium-sized effect of condition (positive vs. negative), $t(247.32) = 5.17$, $p < .001$, $BF_{10} = 4.91 \times 10^{4}$, Cohen's $d = 0.62$, demonstrating the sensitivity of causally informed implicit evaluations to the valence of outcomes. We then investigated the difference in implicit evaluations between causal and noncausal stimuli separately in each condition. In the positive condition, implicit evaluations of causal stimuli were found to be significantly more positive than implicit evaluations of co-occurring (inert) stimuli, $t(167) = 6.69$, $p < .001$, $BF_{10} = 2.66 \times 10^{7}$, Cohen's $d = 0.52$, replicating the result obtained in Study 3. In the negative condition, although the effect was descriptively in the theoretically expected direction, implicit evaluations of causal stimuli and co-occurring (inert) stimuli were found not to significantly differ from each other, $t(126) = -1.46$, $p = .146$, $BF_{01} = 3.59$, Cohen's $d = 0.13$.

5.3. Discussion

Study 4 demonstrates that implicit evaluations of causal objects track the valence of the outcome that they cause. In line with Studies 1–3, we found that stimuli causally responsible for positive outcomes are evaluated more positively than stimuli merely co-occurring with such outcomes. In contrast, explicit evaluations of negative causal stimuli were more negative than negative co-occurring stimuli, whereas implicit evaluations of both types of stimuli were similar. These findings indicate that "causal objects" are not always evaluated more favorably than "non-causal objects"; rather, objects inherit the specific valence of the outcomes they cause. Unexpectedly, evaluative differences between causal and merely co-occurring stimuli were smaller in the negative than in the positive domain, a finding to which we return in the General Discussion.

6. Study 5

In Studies 1–4 we obtained differences in implicit evaluation between objects that caused valenced outcomes and objects that merely co-occurred with those outcomes. This difference is consistent with the possibility that implicit evaluations are genuinely sensitive to causal status. However, notably, these results may also be attributed to preferential attention to or deeper processing of causal over merely statistical relationships (see also Hughes et al., 2018). In the present study, we seek to eliminate such alternative explanations by relying on the idea that causal relationships are inherently directional (Lagnado & Sloman, 2006). Specifically, we investigated implicit evaluations of objects that caused the appearance of a positive stimulus as compared to those that were shown to be caused by the same positive stimulus. We hypothesized that an object that caused the appearance of a positive stimulus would be evaluated more positively than one caused by the appearance of a positive stimulus. The first object is useful in a way that the second object is not: By intervening on it, we can get something we want. Hence, it would be reasonable for evaluations of the first object to be more positive than those of the second.

6.1. Method

6.1.1. Participants and design

The method of recruitment and exclusion criteria were identical to those used in Studies 1–4. The sample size and exclusion criteria, analytic choices, and predictions were preregistered (https://aspredicted.org/u75xm.pdf). The sample size was determined so as to provide 0.80 power to detect a small effect of Cohen's $d = 0.17$, obtained on the IAT in a pilot study involving 469 participants.[7] Of an initial sample of 1217, 15 participants were excluded for failing to complete the IAT and five for excessively fast responding, resulting in a final sample size of 1197. The sample consisted of 808 women and 378 men; 11 participants did not report their gender. The mean age was 39 years ($SD = 16$ years).

[7] Data for the pilot study are available on OSF (https://osf.io/afwyk/).

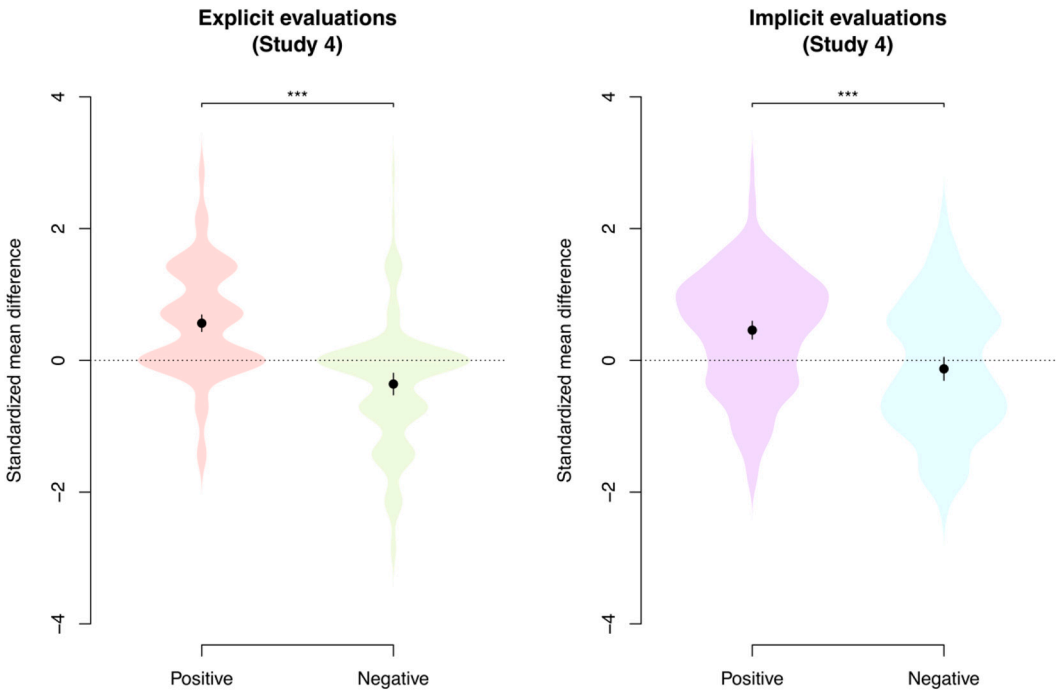


Fig. 4. Distribution of explicit and implicit evaluations (Study 4), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for the causal over the co-occurring (inert) stimulus. Differences between conditions significant at $p < .001$ are marked ***.

Participants were randomly assigned to a cause condition ($n = 649$) or an effect condition ($n = 548$) in a between-subjects design.

6.1.2. *Materials*

The colored shapes, the image of a diamond, and the image of smoke used in previous studies were retained for use in Study 5. However, an image of a different machine was used. The machine was similar to the machine used in Studies 3 and 4, with the crucial exception that it had two, rather than one, emission slots at the bottom (see Table 1). Moreover, an image of coal was newly added to the materials used.

6.1.3. *Procedure and measures*

The procedure and measures were similar to those implemented in Studies 1–4. That is, in a learning phase, participants learned about the operation of a machine, including the difference between causal and co-occurring stimuli followed by a test phase in which participants completed an IAT and explicit measures of memory and evaluation.

At the outset of the learning phase, participants in both conditions were instructed that they would learn about a machine that is designed to emit diamonds but, if not operated properly, breaks apart and emits coal and toxic smoke. Importantly, for participants in the cause condition, the machine was described as a diamond maker. These participants were informed that the machine could operate properly only if an activator of the right color was inserted into the slot at the top. By contrast, for participants in the effect condition, the machine was described as a diamond polisher. These participants were informed that the machine could operate properly only if diamond, rather than coal, was inserted into the slot at the top. Participants in both conditions were told that, along with the diamond or coal, a colored shape would also be emitted by the machine: in the cause condition, the activator that was originally inserted, and in the effect condition, a byproduct.

Following the initial instructions, participants in both conditions were exposed to a total of 16 trials involving the machine and two of the four sets of colored shapes. For each participant, one color was randomly assigned to serve as the color co-occurring with positive events and another color was randomly assigned to serve as the color co-occurring with negative events. Of the 16 trials, eight trials were positive trials and eight trials were negative trials in both conditions. The order of positive and negative trials was individually randomized.

On positive trials, a proper activator was shown to enter the machine through the slot above, upon which the machine emitted a diamond and a shape of a certain color (positive color). On negative trials, an improper activator was shown to enter the machine through the slot above, upon which the machine emitted coal and a shape of a different color (negative color), broke apart, and started smoking.

As mentioned above, the nature of proper and improper activators differed across the cause and effect conditions. Specifically, in the cause condition, shapes of one color served as proper activators and shapes of a different color served as improper activators and were then also shown to exit the machine. By contrast, in the effect condition, the positive stimulus (diamond) and negative stimulus (coal) served as proper and improper activators and colored shapes were shown only to exit, but not to enter, the machine. As such, across the two conditions, shapes of one color consistently co-occurred with positive events and shapes of a different color consistently co-occurred with negative events. However, whereas in the cause condition, shapes of the positive color were causally responsible for positive events and shapes of the negative color were causally responsible for negative events, in the effect condition, shapes of the positive color were caused by positive events and shapes of the negative color were caused by negative events.

The test phase was similar to the test phase of Studies 1–4, with two exceptions. First, the IAT measured implicit evaluations of positive versus negative, rather than causal versus co-occurring, stimuli. Second, participants completed only two explicit evaluation items: one for the color co-occurring with positive events and one for the color co-occurring with negative events.

6.2. *Results*

The distribution of explicit and implicit evaluations is shown in Fig. 5. As expected, evaluations were, on average, positive in both conditions, reflecting a preference for the color co-occurring with positive events

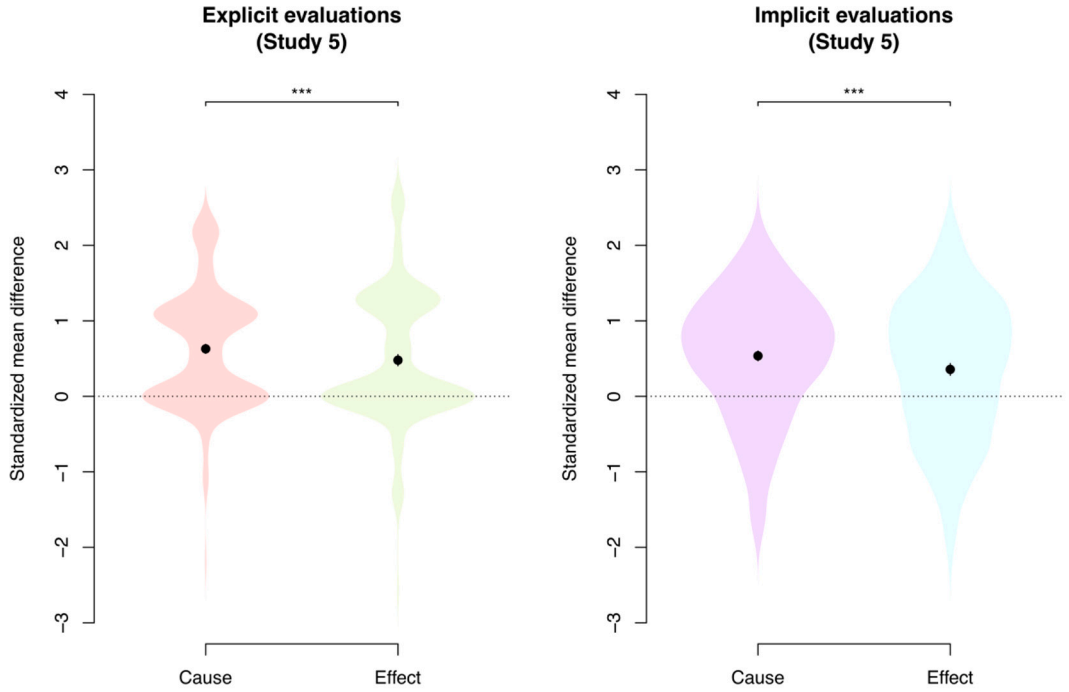


Fig. 5. Distribution of explicit and implicit evaluations (Study 5), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for the stimulus co-occurring with positive rather than negative events. Differences between conditions significant at $p < .001$ are marked ***.

over the color co-occurring with negative events. However, descriptively, it appears that the causal status of the colored stimuli modulated the strength of both explicit and implicit evaluations. Specifically, explicit evaluations seemed to be more positive in the cause condition ($M = 1.15$, $SD = 1.43$) in which the colored stimuli were causally responsible for positive and negative events than in the effect condition ($M = 0.75$, $SD = 1.37$) in which the same colored stimuli were the consequence of positive and negative events. A similar trend was apparent on the implicit measure, with a stronger preference in favor of positive stimuli over negative stimuli emerging in the cause condition ($M = 0.34$, $SD = 0.54$) than in the effect condition ($M = 0.22$, $SD = 0.58$).

6.2.1. Explicit evaluations

As predicted, we found a significant effect of valence on explicit evaluations, $t(1186) = 23.58$, $p < .001$, $BF_{10} = 2.20 \times 10^{97}$, Cohen's $d = 0.68$, indicating that stimuli co-occurring with positive events were generally evaluated more positively than stimuli co-occurring with negative events across both conditions. Second, also as predicted, we obtained strong evidence that causal status modulated the strength of explicit evaluations, with a larger difference emerging in the cause condition than in the effect condition, $t(1165.5) = 5.02$, $p < .001$, $BF_{10} = 1.37 \times 10^{4}$, Cohen's $d = 0.29$.

6.2.2. Implicit evaluations

As predicted, we found a significant effect of valence on implicit evaluations, $t(1196) = 17.67$, $p < .001$, $BF_{10} = 4.30 \times 10^{58}$, Cohen's $d = 0.51$. Second, also as predicted, we obtained strong evidence that the difference in evaluation between positive and negative stimuli was larger in the cause condition than in the effect condition, $t(1122.3) = 3.59$, $p < .001$, $BF_{10} = 40.82$, Cohen's $d = 0.21$.

6.3. Discussion

Study 5 provides further evidence that causal status influences implicit evaluation. Specifically, in a preregistered between-participant design, the difference in implicit (and explicit) evaluations between stimuli co-occurring with positive events and stimuli co-occurring with negative events was found to be larger when those stimuli were shown to be causing, rather than caused by, those events. Crucially, in this design, both stimuli participated in a causal relationship. As such, the results cannot be attributed to heightened attention to, or deeper processing of, causal over noncausal relationships. Rather, it seems that implicit evaluations are genuinely sensitive to the distinction between causes and effects.

7. Study 6

Studies 1–5 have provided strong evidence for the sensitivity of implicit evaluation to causal status; however, these studies, along with relevant past work, leave open the question of the mental representations subserving the effect. According to available theoretical perspectives, there are at least two possibilities. On the one hand, implicit evaluation may reflect causal status because when completing the IAT, participants consult a causal model of the environment, or propositional representations characterizing that model, online (the "causal computation" view). Alternatively, participants may use causal knowledge to precompile compressed value representations during learning, and then query these precompiled representations automatically when completing the implicit measure (the "causal access" view).

As a test of these two competing possibilities, in Study 6 we implemented a revaluation paradigm (Adams & Dickinson, 1981; Daw et al., 2011; Gershman et al., 2014; Kurdi et al., 2019). Specifically, participants in all three conditions of the experiment observed a causal system

in which A caused B; B initially caused a negative outcome (toxic smoke); A was correlated with C; and C did not cause B. For example, blue shapes (A) may have been shown to cause the appearance of orange tokens (B). These orange tokens may then have entered a machine and caused it to break and to emit toxic smoke. Green shapes (C) may have consistently co-occurred with blue shapes (A) do but did not cause the appearance of orange tokens (B). In the baseline condition, the information described above was the only training information to which participants were exposed.

In the theoretically crucial revaluation and relearning conditions (but not in the baseline condition described above), the first part of the learning phase was followed by a second part. During this second part, the negative outcome (toxic smoke) was replaced with a positive outcome (the machine dispensing diamonds).[8] In the revaluation condition, participants were exposed only to orange tokens (B) causing the appearance of diamonds; in the relearning condition, participants observed the entire new causal sequence with the valence of the outcome reversed (A → B → DIAMOND).

If implicit evaluation has the ability to consult the updated causal model online, as the causal computation view predicts, then implicit preference for blue shapes (A) over green shapes (C) should increase in both conditions. That is, participants would use the new information in their updated causal model to adjust their implicit evaluations. However, if (in line with the causal access view) implicit evaluation reflects representations that encode only the evaluative implications of the initial causal model in a compressed format, it is possible that implicit evaluations will be updated in the relearning but not in the revaluation condition. The reason for this difference is that updating the evaluation of A vs. C in the revaluation condition requires productively combining two pieces of information never shown together based on causal principles: namely that blue shapes (A), but not green shapes (C), cause orange shapes (B) and that orange shapes (B) newly cause a positive outcome (diamonds). By contrast, updating the evaluation of A does not require any such recombination in the relearning condition; it is made visually apparent to participants.

7.1. Method

7.1.1. Participants and design

The method of recruitment and exclusion criteria were identical to those used in Studies 1–5. The sample size and exclusion criteria, analytic choices, and predictions were preregistered (https://aspredicted.org/ts9bp.pdf). Some non-preregistered Bayesian analyses were also conducted using theoretically derived ordinal constraints on condition means (Haaf & Rouder, 2017; Rouder, Haaf, & Aust, 2018). These analyses are marked as exploratory below. Of an initial sample of 1789 (recruited in two batches in line with the preregistration), 19 participants were excluded for failing to complete the IAT and 13 for excessively fast responding, resulting in a final sample size of 1757. The sample consisted of 1197 women, 540 men, and 9 participants with other gender identities; 11 participants did not report their gender. The mean age was 39 years (SD = 16 years). Participants were randomly assigned to a baseline condition (n = 614), a revaluation condition (n = 569), or a relearning condition (n = 574) in a between-subjects design.

7.1.2. Materials

The image of a diamond and the image of smoke used in previous studies and the image of the machine used in Studies 3 and 4 were retained for use in Study 6. The colored shapes used in previous studies were replaced by a new set of five shapes (a circle, a square, a triangle, a hexagon, and a star), each created in five different colors (blue, green, orange, violet, and yellow) selected to make sure that colorblind participants were able to reliably distinguish between any two colors. Moreover, a new image of a smaller machine was added to the stimuli (see Table 1).

7.1.3. Procedure and measures

The procedure and measures were similar to those implemented in Studies 1–5. That is, participants learned about the operation of two machines, including the difference between causal and co-occurring stimuli, in a learning phase, and completed an IAT, followed by explicit measures of memory and evaluation, in a test phase.

The content of the learning phase differed depending on participants' condition assignment. In the first part of the learning phase, participants in all three conditions (baseline, revaluation, and relearning) completed the same learning experience, which was followed by a second learning experience for participants assigned to the revaluation and relearning conditions.

The first part of the learning phase in all three conditions (which constituted the only learning phase in the baseline condition) was modeled after the learning phase in Study 4, with some important modifications. First, participants were informed about the intended operation of two machines. Specifically, they were told that the purpose of the two machines was to make diamonds. To be activated, a shape of a certain color (causal color) had to touch the activator surface of the first machine; if an activator of a different color tethered to the activator of the first color (co-occurring color) touched this surface, the machine would remain inert. When successfully activated, the first machine would dispense a token of a third color (token color), which would enter the second machine. The second machine would then dispense a diamond. However, participants were also told that the second machine may occasionally malfunction; in these cases, it would break apart and emit toxic smoke. Crucially, although the verbal instructions did not alert participants to this fact, the first part of the learning phase included only trials with the negative outcome (i.e., the second machine malfunctioning) and no trials with the positive outcome (i.e., the second machine dispensing diamonds).

Following these instructions, participants were exposed to a total of 12 trials involving the two machines and three of the five sets of colored shapes. For each participant, one color was randomly assigned to serve as the color ultimately causally responsible for the negative outcome (causal color), another color to serve as an inert color that consistently co-occurred with the causal color but played no causal role in bringing about this outcome (co-occurring color), and a third color to serve as the proximate cause of the outcome (token color). Of the 12 trials, eight trials were causal trials and four trials were inert trials. The first two trials were always causal trials, whereas the order of the remaining trials was individually randomized.

Causal stimuli and co-occurring (inert) stimuli were always shown to be tethered to each other. On causal trials, a randomly selected causal stimulus was shown to drop onto the activator surface of the first machine, upon which the light bulb switched on, indicating that the machine was operational, and the machine emitted a token stimulus. A randomly selected co-occurring (inert) stimulus was shown to be connected to the positive causal stimulus but did not touch the activator surface of the machine. The token stimulus then entered the second machine, which broke apart and emitted toxic smoke. Inert trials involved the same stimuli as causal trials; however, instead of the causal stimulus, the co-occurring (inert) stimulus touched the activator surface of the first machine and the machine was not activated.

Whereas participants in the baseline condition underwent only the learning experience described above, in the revaluation and relearning conditions, the first part of the learning phase was followed by a second part. After the first part of the learning phase, participants in the

[8] The negative-to-positive rather than positive-to-negative order was selected in this study because in Study 5 we found that causal status has a bigger effect on implicit (and explicit) evaluations in the positive rather than the negative domain. As such, we anticipated that any updating as a result of causal status would be more likely when the valence of the outcome is positive rather than negative.

revaluation condition were informed that the second machine was now fixed and worked properly. More precisely, participants were told that, as intended, if a shape of the token color entered the second machine, the second machine would now produce a diamond. Then they observed 12 revaluation trials on which, as previously described, a token stimulus entered the second machine, upon which the second machine emitted a diamond. Crucially, the causal and co-occurring (inert) stimuli were neither mentioned in the instructions nor shown in the second part of the learning phase. However, participants were told that they should assume that the first machine (although not shown anymore) still functioned the same way as before.

Participants in the relearning condition received instructions similar to participants in the revaluation condition, but they were also additionally reminded that the first machine worked only if the causal color was put on top of the activator. Then they observed twelve trials of which 8 were positive causal trials and four were inert trials.[9] On positive causal trials, the causal color activated the first machine, the first machine dispensed a token, the token entered the second machine, and the second machine dispensed a diamond. As such, similar to the revaluation condition, the valence of the outcome was reversed, but unlike in the revaluation condition, participants were able to observe the entire new causal sequence. The inert trials were identical to the inert trials shown during the first part of the learning phase.

The test phase was similar to the test phase of Studies 1–5, with three exceptions. First, participants responded to a manipulation check item before the IAT.[10] Second, the words "good", "great," "fantastic", "pleasant," and "wonderful" were used as positive attribute stimuli and the words "awful," "bad," "horrible," "terrible," and "unpleasant" were used as negative attribute stimuli on the IAT. Third, participants completed explicit evaluation items for the causal color, the co-occurring color, and the token color on 100-point slider scales whose left endpoint, midpoint, and right endpoint were labeled "extremely negative," "neither negative nor positive," and "extremely positive," respectively.

7.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 6. In line with the preregistered prediction on the basis of a pilot study, participants exhibited positive explicit ($M = 9.98$, $SD = 29.89$) and implicit ($M = 0.15$, $SD = 0.50$) evaluations in the baseline condition indicating a preference for the causal over the co-occurring stimulus. This preference most likely emerged as a result of the initial verbal instructions, which described the intended operation of the machines. Although participants in the baseline condition were exposed only to negative trials, these verbal instructions had indicated that the intended role of the causal color was to bring about a positive outcome, whereas the co-occurring color was causally inert. Given our focus on the comparison between the baseline and the revaluation and relearning conditions, this expected but theoretically inconsequential result will not be discussed further.

As predicted, the explicit preference observed in the baseline condition became stronger in both the revaluation ($M = 15.61$, $SD = 38.28$) and the relearning conditions ($M = 20.59$, $SD = 37.27$). Crucially, in line with the causal access hypothesis, implicit evaluations shifted in the positive direction in the relearning condition ($M = 0.21$, $SD = 0.52$) but remained equivalent to baseline in the revaluation condition ($M = 0.14$, $SD = 0.51$).

7.2.1. Explicit evaluations

Explicit evaluations significantly differed from each other across the three conditions, $F(2, 1733) = 13.36$, $p < .001$. Explicit evaluations were significantly different from zero in the baseline condition, indicating a preference for the causal over the co-occurring color, $t(1733) = 6.98$, $p < .001$, Cohen's $d = 0.33$. This difference in favor of the causal over the co-occurring color became significantly larger in both the revaluation, $t(1733) = 2.73$, $p = .006$, Cohen's $d = 0.16$, and the relearning conditions, $t(1733) = 5.16$, $p < .001$, Cohen's $d = 0.32$. As such, for explicit evaluations, all preregistered predictions were confirmed. Finally, in an exploratory analysis, we probed whether the relearning and revaluation conditions differed from each other and found significantly stronger updating in the former than in the latter, $t(1733) = 2.37$, $p = .018$, Cohen's $d = 0.13$.

7.2.2. Implicit evaluations

Implicit evaluations significantly differed from each other across the three conditions, $F(2, 1754) = 3.15$, $p = .043$. Implicit evaluations were significantly different from zero in the baseline condition, indicating a preference for the causal over the co-occurring color, $t(1754) = 7.15$, $p < .001$, Cohen's $d = 0.30$. This difference in favor of the causal over the co-occurring color did not change in the revaluation condition, $t(1754) = -0.28$, $p = .778$, Cohen's $d = 0.01$, but it became significantly stronger in the relearning condition, $t(1754) = 2.05$, $p = .040$, Cohen's $d = 0.12$. As such, for implicit evaluations, all preregistered predictions in line with the causal access hypothesis were confirmed. Finally, in an exploratory analysis, we probed whether the relearning and revaluation conditions differed from each other and found significantly stronger updating in the former than in the latter, $t(1754) = 2.29$, $p = .022$, Cohen's $d = 0.13$.

7.2.3. Exploratory Bayesian analyses

We also conducted exploratory Bayesian analyses to compare the relative plausibility of three hypotheses in light of the data. A strong version of the causal computation hypothesis would predict that responding in the revaluation and relearning conditions should be equivalent and explicit evaluations in both of these conditions should be more positive than in the baseline condition. A weak version of this hypothesis predicts that both the revaluation and the relearning condition should be more positive than the baseline condition; however, these two conditions may not be equivalent to each other. Finally, the causal access hypothesis predicts that the baseline and revaluation conditions should be equivalent to each other, and the relearning condition should be more positive than both.

With explicit evaluations as the dependent measure, the weak version of the causal computation hypothesis was found to provide the best description of the data. This hypothesis clearly outperformed the causal access hypothesis, BF = 4.22, and to some degree also the strong version of the causal computation hypothesis, BF = 1.73. In addition, the weak causal computation hypothesis also outperformed the null hypothesis (under which all three condition means are predicted to be equal), BF = 3088.98, and the full model (under which any pattern of condition means is permissible), BF = $6 \in [0, 6]$.

However, with implicit evaluations as the dependent measure, exploratory Bayesian analyses indicate that the causal access hypothesis provides the best fit with the data. This hypothesis outperformed both the strong, BF = 15.32, and the weak versions of the causal computation hypothesis, BF = 21.47. In addition, the causal access hypothesis also (weakly) outperformed the null hypothesis, BF = 1.24, and convincingly outperformed the full model BF = $5.96 \in [0, 6]$.

[9] Given that we sought to keep the total number of trials constant across the revaluation and relearning conditions, the revaluation condition contained more causal trials (12) than the relearning condition did (8). Based on this difference, the revaluation condition could have been expected to spuriously produce stronger learning than the relearning condition. However, as explained below, we obtained the opposite pattern of results on both the explicit and the implicit measure. As such, this incidental feature of the design does not undermine our conclusions. We thank an anonymous reviewer for alerting us to this issue.

[10] Results on the manipulation check item, which were in line with theoretical predictions, are available on OSF (https://osf.io/afwyk/).

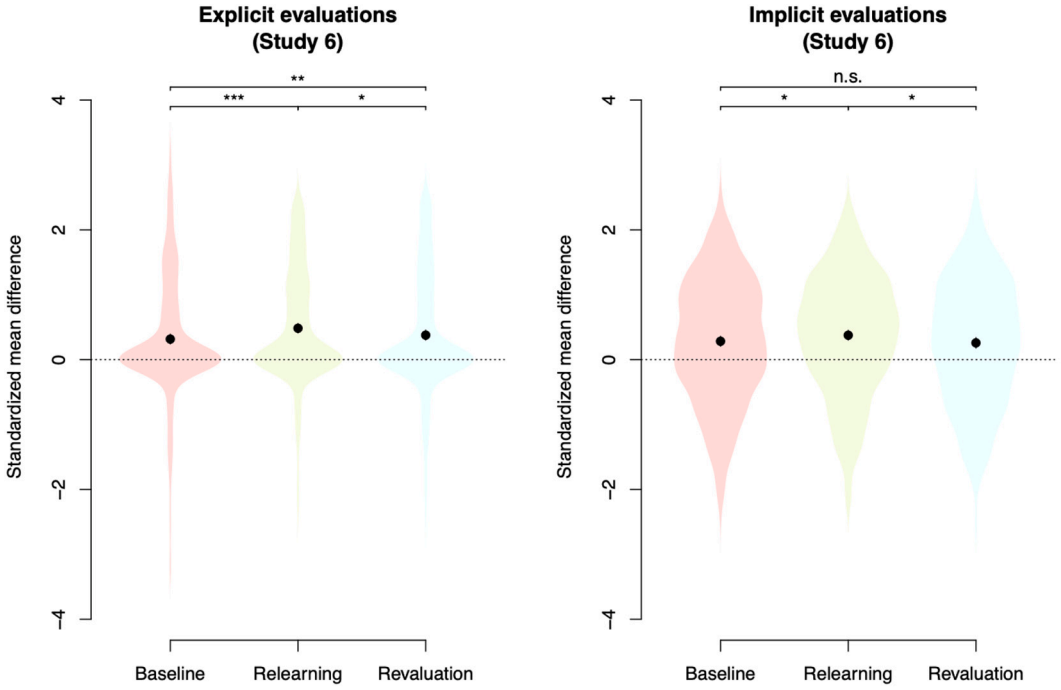


Fig. 6. Distribution of explicit and implicit evaluations (Study 6), displayed in standardized units to ensure comparability. The dashed line shows neutrality and the solid dot shows the mean, with error bars corresponding to 95% confidence intervals. Positive scores indicate preference for causal over the co-occurring (inert) stimulus. Differences between conditions significant at $p < .05$ are marked *, at $p < .01$ are marked **, and at $p < .001$ are marked ***; n.s. = not significant.

7.3. Discussion

In Study 6, we probed the nature of the mental representations underlying the sensitivity of implicit evaluations to causal information. Specifically, in a baseline condition, participants learned the causal sequence A → B → SMOKE, along with the fact C is perfectly correlated with A but does not cause B. Subsequently, the initially negative outcome (smoke) was replaced with a positive outcome (diamonds). In the revaluation condition, participants were exposed only to the new causal relation B → DIAMONDS, whereas in the relearning condition, the entire new causal sequence was shown (A → B → DIAMONDS).

Explicit evaluations were significantly updated in the expected direction in both cases, suggesting that they are supported by flexible online computations over a causal model or the propositional representations encoding the details of such a model. By contrast, implicit evaluations shifted only in the relearning but not in the revaluation condition, thus providing initial evidence for the causal access hypothesis. The latter result suggests that implicit evaluation may emerge from relatively compressed value representations precompiled in a causally informed manner during initial learning rather than from online computations over a causal model.

8. General discussion

We conducted six experiments to probe the sensitivity of implicit evaluations to causal structure and to explore the cognitive mechanisms underlying such sensitivity. In Studies 1A and 1B, implicit evaluations of causal stimuli were found to be more positive than implicit evaluations of co-occurring stimuli, although co-occurring stimuli appeared closer to inherently positive stimuli in both space and time. In Study 2, we replicated the same result in a design relying exclusively on observational learning in the absence of any verbal instructions on the operation of the machine. In Study 3, we eliminated two possible confounds inherent to the design of Studies 1 and 2: temporal primacy (Carney & Banaji, 2012) and attentional capture by valenced stimuli (Theeuwes & Belopolsky, 2012). In Study 4 we showed that causal objects inherit the valence of the events that they cause. And, in Study 5, we demonstrated that objects that cause the appearance of positive stimuli are evaluated more positively on implicit measures than objects caused by the appearance of positive stimuli. As such, the results of Study 5 make it unlikely that the results of Studies 1–4 had emerged merely due to increased attention to or deeper processing of causal relative to mere statistical relationships.

The most parsimonious explanation for this set of results is that implicit evaluation is not exclusively sensitive to statistical relatedness but is instead also responsive to differences in causal status. This finding is difficult to reconcile with those dual-process theories in psychology and philosophy that posit the existence of an implicit system that represents co-occurrence information alone and is fully impervious to causal relationships (e.g., Gendler, 2008; Lieberman et al., 2002; Madva, 2016; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Neither are the present findings easily accounted for by augmented versions of associative learning, such as the Rescorla–Wagner model (Rescorla & Wagner, 1972), given that causal and co-occurring stimuli in Studies 3 and 4 did not differ from each other in terms of their temporal or statistical relationship with positive and negative outcomes, either within or across trials.

In contrast, the sensitivity of implicit evaluations to causal structure is consistent with theories that posit a role for causal information in implicit evaluation (e.g., De Houwer, 2018; Kurdi & Dunham, 2020; Mandelbaum, 2016; Sloman, 2014). As such, the present findings accord with prior research showing that implicit evaluations of individuals who cause bad (or good) events are more negative (or positive) than implicit evaluations of individuals who merely co-occur with such events (e.g., Hughes et al., 2018). Moreover, they extend this line of work by systematically demonstrating a role of causal structure in implicit evaluations of nonsocial targets, and by providing evidence for the idea that implicit evaluation is sensitive to the distinction between causes and effects rather than simply assigning a general evaluative bonus to any stimulus participating in a causal relationship. Additionally, in extending work by Gawronski and colleagues (Gawronski & Brannon, 2021; Heycke & Gawronski, 2020; Hu et al., 2017), the current results show

that implicit evaluations can reflect causal status learned under conditions that do not involve language as a privileged source of causal information. Notably, such effects seem to emerge even when causal information reflects relatively fine-grained differences, such as the distinction between causes and effects.

Finally, in Study 6, we asked whether implicit evaluations merely access representations generated from causal principles, or whether they spontaneously compute over causal principles. According to recent theoretical perspectives, implicit and explicit evaluation share the ability to perform online computations over causal models (Sloman, 2014), or they operate over the same propositional representations encoding relational information, including causal relationships (De Houwer, 2018; Mandelbaum, 2016). Seemingly contrary to this view, in Study 6, only explicit evaluations could flexibly incorporate new causal information requiring online use of a causal model; implicit evaluations did not.

According to this model participants might, of course, have precompiled the novel causal relation implied by the stimuli that they had encountered offline, prior to completing the IAT. Had they done so, this causal relation could have been reflected by their IAT performance. This kind of spontaneous precompilation has been observed in other experimental settings (e.g., Gershman et al., 2014), although its linkage to the IAT, or implicit evaluations more generally, remains to be demonstrated. However, it appears that participants in the present experiment did not spontaneously precompile the novel causal relation, or at least not in a format available to the automatic processes underlying IAT performance.

Naturally, this raises the question: When do people spontaneously precompile the kinds of representations over which automatic processes operate, and when do they not? This question is a very general one: There are many representations that could, in principle, be precompiled but are usually not (e.g., the sum 1234 + 945 =?), whereas there are others that generally are (e.g., the words one will use to propose to a would-be spouse). As such, we leave it for future work to develop an overarching theory of when and how "offline" cognitive resources are invested to precompile representations in a way that is useful for future thought and action (see e.g., Icard, Cushman, & Knobe, 2018; Mattar & Daw, 2018). More specifically, it remains to be seen whether and under what conditions such processes of "representational exchange" affect the representations that are available to automatic, implicit processes.

Another important avenue for future research is to assess whether automatic processes have the general capacity to compute over causal principles although this capacity was not invoked, or its effect was not detected, in Study 6. We cannot rule out these possibilities decisively, but regard each as implausible. First, it is conceivable that automatic processes can compute causal relations in general but did not do so in Study 6 specifically. This is unlikely for three reasons: The causal inference was easy to make; it was highly task-relevant; and participants readily performed it when responding to the explicit measure of evaluation (and the manipulation check completed before the IAT). Second, it is possible that automatic processes did compute a novel relation over causal principles in Study 6, but that we did not detect it. For instance, perhaps novel relations available to automatic processes sometimes fail to "overwrite" preexisting relations, as has been suggested in prior research using the IAT (e.g., Gregg, Seibt, & Banaji, 2006). However, the relearning condition in Study 6 illustrates precisely such a case of "overwriting", but in a setting where the novel causal relation need not be computed but is instead visually manifest during learning. Still, future work should provide more stringent and extensive tests of the hypothesis that automatic processes can access, but not compute, novel relations derived from causal principles.

These limitations notwithstanding, the present findings expand on recent results by Kurdi et al. (2019) who also used the distinction between model-based online computation and precompiled value representation to examine the representational structures underlying implicit evaluation. Using a paradigm similar to the one implemented in Study 6 of the present paper, Kurdi and colleagues found that causal information can be used to revise implicit evaluations only to the extent that participants experience the entirety of a new causal chain and not when updating would require productively combining multiple parts of a new causal model online. However, crucially, in the experiments by Kurdi et al. (2019), the initial difference in evaluation between the two targets was due to a difference in the valence of their consequences. By contrast, in the present Study 6, the focal stimuli consistently co-occurred with outcomes of the same valence but differed in causal status.

An element of the present findings from Study 6 also dovetails with prior work by Fernbach and Sloman (2009). Specifically, in this study we found that implicit evaluations were sensitive to relearning involving the full causal chain but not to a revaluation manipulation that would have required productively recombining two pieces of information learned on separate occasions over. Although the experiments conducted by Fernbach and Sloman differ in their procedure, they also suggest that participants have an easier time making accurate causal inferences in situations where all relevant information is presented within the same trial (e.g., A → B → DIAMONDS) than in situations in which the same information is presented across different sets of trials (e.g., A → B and B → DIAMONDS). Crucially, however, Fernbach and Sloman (2009) used only explicit measures of causal judgment. By contrast, the present results show that implicit evaluations can remain insensitive to the second type of learning even when explicit evaluations (and explicit reports of causal structure) are sensitive to it.

More generally, Kurdi and Dunham (2020) have recently suggested that implicit evaluations should be amenable to updating to the extent that updating does not require access to high-dimensional details of the original learning experience, such as the original causal model in the present studies. This view shares an important commonality with propositional theories (e.g., De Houwer, 2018; Mandelbaum, 2016) in that it recognizes the possibility that implicit evaluation, much like explicit evaluation, can be sensitive to a wide range of inputs beyond mere co-occurrence, including causal relationships, as well as instructions about stimulus relations (De Houwer, 2006), mental state reasoning (Kurdi et al., 2020), and normatively accurate and inaccurate applications of propositional logic (Kurdi & Dunham, 2021).

However, in opposition to propositional theories, Kurdi and Dunham (2020) have suggested that explicit and implicit evaluation may differ from each other in the format of the representations on which they rely. Specifically, the former are thought to reflect high-dimensional details of past learning, whereas the latter are thought to emerge from compressed, low-dimensional representations of the same information. As such, implicit processes may enable swift automatic responding to stimuli encountered in the environment while at the same time retaining some (albeit limited) ability to draw on representations of complex relationships among such stimuli. The present studies offer initial support for this view.

The proposal by Kurdi and Dunham (2020) shares important commonalities not only with propositional accounts of implicit evaluation but also with the associative–propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2018). The APE model posits that implicit measures principally tap associative representations, while explicit measures typically tap propositional representations. From this standpoint, causal relationships might not be predicted to influence patterns of responding on the IAT. However, importantly, the APE model allows for relational information to influence implicit evaluations under a limited set of conditions. However, currently, it cannot generate strong predictions about when this will occur. From the perspective of the APE model, the present findings illustrate an important influence of this kind and thus help identify the set of conditions under which implicit evaluations can show sensitivity to relational information. Alternatively, from the perspective of the proposal by Kurdi and Dunham (2020), they confirm a central, a priori prediction that implicit and explicit evaluations should both be sensitive to relational information (Studies 1–5) but that the former should rely on memory structures that

represent this information in a cached, more compressed format (Study 6).

In addition to these main findings, an initially unexpected result emerged from Studies 4 and 6: The difference in both explicit and implicit evaluations between causal and noncausal stimuli was attenuated when the outcome was negative rather than positive. While we are unable to conclusively explain this result here, we can speculate about some alternatives that may be tested in future work. The result may be specific to causal learning. Specifically, causal stimuli (perhaps given their greater potency than merely predictive stimuli) may by default be associated with positivity. Such positivity may then be reduced as a result of an opposing process when outcomes are negative. Alternatively, as a result of negativity dominance (Rozin & Royzman, 2001), information about negative events may simply drown out many other forms of information, including information about causality and other forms of relational information. In line with the latter possibility, a recent meta-analysis has found that both explicit and implicit evaluations tend to be more sensitive to relational information in the positive than in the negative domain (Kurdi, Morehouse, & Dunham, 2022).

The asymmetry between positive and negative outcomes adds to the complexity of the present findings, but it does not challenge our interpretation of them. Across all studies reported in the present paper, the results allow for two parsimonious explanations. First, the results may be accounted for by two additive effects: as hypothesized, enhanced sensitivity to valence for casual (vs. merely associated) stimuli and, unexpectedly, enhanced positivity for causally efficacious (vs. causally inefficacious) stimuli. Second, the results may be accounted for by an interaction, such that implicit evaluations are especially sensitive to factors that cause (vs. predict) positive outcomes, but are not especially sensitive to factors that cause (vs. predict) negative outcomes. In either case, it is clear that implicit evaluation can be sensitive to the difference between causal versus merely predictive relations.

In the present project, we chose to examine several variations to the same paradigm to be able to generate inferences about that paradigm with a high degree of internal validity. We hope that future work will expand on the current studies in multiple ways. First, given that the present experiments all relied on the same basic procedure, our argument should be understood as a proof of concept about the ability of implicit evaluation to reflect causal information; the generality and boundary conditions of this effect remain to be investigated. For instance, it is unclear whether implicit evaluations would be sensitive to causal information under incidental encoding conditions or encoding conditions involving cognitive load (see Fan, Bodenhausen, & Lee, 2021).

Second, to ensure the comparability of results, all of the present experiments relied on the same implicit measure. However, different implicit measures are characterized by subtly different modes of operation (De Houwer & Moors, 2010). As such, it may be instructive to replicate the present results using implicit measures other than the Implicit Association Test.

Third, although the effects of causal information on implicit (and explicit) evaluation were robust in the present experiments, they tended to be quite small in size. Related to the more general point about boundary conditions above, future work should investigate whether these effects are generally small, or whether larger effects may be produced by implementing stronger manipulations or relying on paradigms different from the one used here.

9. Conclusion

Across six studies, we provide evidence against the idea that implicit evaluation reflects exclusively co-occurrence information and is impervious to causal structure. This may be how we seem to know immediately and intuitively that acting dour and buying diapers won't increase a couple's chances of conception. However, such knowledge seems to be activated immediately and intuitively not because implicit evaluation is mediated by online consultation of a causal model but rather because it can query precompiled representations derived from a causal model offline.

CRediT authorship contribution statement

Benedek Kurdi: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft. **Adam Morris:** Conceptualization, Methodology, Writing – review & editing. **Fiery A. Cushman:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of Competing Interest

Benedek Kurdi is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition.

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. Sec. B, 33*(2b), 109–121. https://doi.org/10.1080/14640748108400816

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognit. Emot., 34*(1), 128–143. https://doi.org/10.1080/02699931.2019.1604322

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task. *Exp. Psychol., 56*(5), 329–343. https://doi.org/10.1027/1618-3169.56.5.329

Blanchflower, D. G. (2009). International evidence on well-being. In A. B. Krueger (Ed.), *Measuring the subjective well-being of nations: National accounts of time use and well-being* (pp. 155–226). Chicago, IL: University of Chicago Press.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci., 1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011

Carney, D. R., & Banaji, M. R. (2012). First is best. *PLoS One, 7*(6), Article e35088. https://doi.org/10.1371/journal.pone.0035088

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

De Houwer, J. (2006). Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learn. Motiv., 37*(2), 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

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

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Soc. Psychol. Bull., 13*(3). https://doi.org/10.5964/spb.v13i3.28046, 49–21.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–196). New York, NY: Guilford Press.

De Houwer, J., & Vandorpe, S. (2010). Using the implicit association test as a measure of causal learning does not eliminate effects of rule learning. *Exp. Psychol., 57*(1), 61–67. https://doi.org/10.1027/1618-3169/a000008

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nat. Neurosci., 18*(5), 767–772. https://doi.org/10.1038/nn.3981

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol., 22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

Fan, X., Bodenhausen, G. V., & Lee, A. Y. (2021). Acquiring favorable attitudes based on aversive affective cues: Examining the spontaneity and efficiency of propositional evaluative conditioning. *J. Exp. Soc. Psychol., 95*, Article 104139. https://doi.org/10.1016/j.jesp.2021.104139

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A *bona fide* pipeline? *J. Pers. Soc. Psychol., 69*(6), 1013–1027. https://doi.org/10.1037//0022-3514.69.6.1013

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

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *J. Exp. Psychol. Learn. Mem. Cogn., 35*(3), 678–693. https://doi.org/10.1037/a0014928

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

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative–propositional evaluation model. *Soc. Psychol. Bull., 13*(3), Article e28024. https://doi.org/10.5964/spb.v13i3.28024

Gawronski, B., & Brannon, S. M. (2021). Attitudinal effects of stimulus co-occurrence and stimulus relations: Range and limits of intentional control. *Personal. Soc. Psychol. Bull., 47*(12), 1654–1667. https://doi.org/10.1177/0146167220982906

Gendler, T. S. (2008). Alief and belief. *J. Philos., 105*(10), 634–663. https://doi.org/10.5840/jphil20081051025

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *J. Exp. Psychol. Gen., 143*(1), 182–194. https://doi.org/10.1037/a0030844

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev., 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. *J. Pers. Soc. Psychol., 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. *J. Pers. Soc. Psychol., 85*(2), 197–216. https://doi.org/10.1037//0022-3514.85.2.197

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *J. Pers. Soc. Psychol., 90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychol. Methods, 22*(4), 779–798. https://doi.org/10.1037/met0000156

Heycke, T., & Gawronski, B. (2020). Co-occurrence and relational information in evaluative learning: A multinomial modeling approach. *J. Exp. Psychol. Gen., 149*(1), 104–124. https://doi.org/10.1037/xge0000620

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. the effects of co-occurrence and relational information on implicit and explicit evaluations. *Personal. Soc. Psychol. Bull., 43*(1), 17–32. https://doi.org/10.1177/0146167216673351

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2018). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personal. Soc. Psychol. Bull., 45*(2), 196–208. https://doi.org/10.1177/0146167218781340

Icard, T., Cushman, F. A., & Knobe, J. (2018). On the instrumental value of hypothetical and counterfactual thought. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 517–522). Austin, TX: Cognitive Science Society.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kurdi, B., & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Soc. Cogn., 38*(Supplement), s42–s67. https://doi.org/10.1521/soco.2020.38.supp.s42

Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition, 214*, Article 104792. https://doi.org/10.1016/j.cognition.2021.104792

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proc. Natl. Acad. Sci., 116*(13), 6035–6044. https://doi.org/10.1073/pnas.1820238116

Kurdi, B., Krosch, A. R., & Ferguson, M. J. (2020). Implicit evaluations of moral agents reflect intent and outcome. *J. Exp. Soc. Psychol., 90*. https://doi.org/10.1016/j.jesp.2020.103990, 103990–12.

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2022). *How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information* (Manuscript submitted for publication).

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *J. Exp. Psychol. Learn. Mem. Cogn., 32*(3), 451–460. https://doi.org/10.1037/0278-7393.32.3.451

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). New York, NY: Oxford University Press.

Le Pelley, M. E., Griffiths, O., & Beesley, T. (2017). Associative accounts of causal cognition. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 13–28). New York, NY: Oxford University Press.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., & Trope, Y. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. In *, Vol. 34. Advances in Experimental Social Psychology* (pp. 199–249). Elsevier. https://doi.org/10.1016/S0065-2601(02)80006-5.

Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese, 193*(8), 2659–2684. https://doi.org/10.1007/s11229-015-0874-2

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs, 50*(3), 629–658. https://doi.org/10.1111/nous.12089

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci., 21*(11), 1609–1617. https://doi.org/10.1038/s41593-018-0232-z

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Dev. Psychol., 48*(5), 1215–1228. https://doi.org/10.1037/a0027440

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *J. Exp. Psychol., 90*(2), 227–234. https://doi.org/10.1037/h0031564

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognit. Emot., 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. *J. Exp. Soc. Psychol., 60*(C), 157–162. https://doi.org/10.1016/j.jesp.2015.05.009

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta–cognitive model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Soc. Cogn., 25*(5), 657–686. https://doi.org/10.1521/soco.2007.25.5.657

Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST model. *J. Pers. Soc. Psychol., 90*(1), 21–41. https://doi.org/10.1037/0022-3514.90.1.21

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proc. Natl. Acad. Sci., 114*(18), 4649–4654. https://doi.org/10.1073/pnas.1619717114

Rand, D. G. (2016). Cooperation, fast and slow. *Psychol. Sci., 27*(9), 1192–1206. https://doi.org/10.1177/0956797616654455

Rehder, B. (2009). Causal-based property generalization. *Cogn. Sci., 33*(3), 301–344. https://doi.org/10.1111/j.1551-6709.2009.01015.x

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cogn. Psychol., 72*, 54–107. https://doi.org/10.1016/j.cogpsych.2014.02.002

Rehder, B. (2015). The role of functional form in causal-based categorization. *J. Exp. Psychol. Learn. Mem. Cogn., 41*(3), 670–692. https://doi.org/10.1037/xlm0000048

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Mem. Cogn., 45*(2), 245–260. https://doi.org/10.3758/s13421-016-0662-3

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: Aplleton-Century-Crofts.

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Commun. Monogr., 85*(1), 41–56. https://doi.org/10.1080/03637751.2017.1394581

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personal. Soc. Psychol. Rev., 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull., 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Sloman, S. A. (2014). Two systems of reasoning: An update. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 69–79). New York, NY: Guilford Press.

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

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personal. Soc. Psychol. Rev., 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: Oculomotor capture by rewarding stimuli. *Vis. Res., 74*(C), 80–85. https://doi.org/10.1016/j.visres.2012.07.024

Tolman, E. C. (1949). There is more than one kind of learning. *Psychol. Rev., 56*(3), 144–155. https://doi.org/10.1037/h0055304