Behavior Research Methods (2024) 56:3161–3194
https://doi.org/10.3758/s13428-023-02291-2

ORIGINAL MANUSCRIPT

Testing the automaticity features of the affect misattribution procedure: The roles of awareness and intentionality

Benedek Kurdi[1,2] · David E. Melnikoff[3] · Jason W. Hannay[4] · Arın Korkmaz[1] · Kent M. Lee[3] · Emily Ritchie[1] · Nicholas Surdel[1] · Heidi A. Vuletich[5] · Xin Yang[1] · B. Keith Payne[6] · Melissa J. Ferguson[1]

Accepted: 8 November 2023 / Published online: 29 November 2023

Abstract

The affect misattribution procedure (AMP) is a measure of implicit evaluations, designed to index the automatic retrieval of evaluative knowledge. The AMP effect consists in participants evaluating neutral target stimuli positively when preceded by positive primes and negatively when preceded by negative primes. After multiple prior tests of intentionality, Hughes et al. (Behav Res Methods 55(4):1558–1586, 2023) examined the role of awareness in the AMP and found that AMP effects were larger when participants indicated that their response was influenced by the prime than when they did not. Here we report seven experiments (six preregistered; $N = 2350$) in which we vary the methodological features of the AMP to better understand this awareness effect. In Experiments 1–4, we establish variability in the magnitude of the awareness effect in response to variations in the AMP procedure. By introducing further modifications to the AMP procedure, Experiments 5–7 suggest an alternative explanation of the awareness effect, namely that awareness can be the outcome, rather than the cause, of evaluative congruency between primes and responses: Awareness effects emerged even when awareness could not have contributed to AMP effects, including when participants judged influence awareness for third parties or primes were presented post hoc. Finally, increasing the evaluative strength of the primes increased participants' tendency to misattribute AMP effects to the influence of target stimuli. Together, the present findings suggest that AMP effects can create awareness effects rather than vice versa and support the AMP's construct validity as a measure of unintentional evaluations of which participants are also potentially unaware.

Keywords Affect misattribution procedure · Automaticity · Awareness · Intentionality · Implicit evaluations

Since its inception as a scientific discipline, psychology has grappled with questions surrounding the nature of attitudes,[1] or evaluations of social entities along a positive–negative continuum (Eagly & Chaiken, 1993; Fazio, 2007). Starting in the 1980s, attention started shifting away from an

Preregistrations, materials, data, and analysis scripts are available via the Open Science Framework (https://osf.io/wfksp/).

✉ Benedek Kurdi
kurdi@illinois.edu

[1] Department of Psychology, Yale University, New Haven, CT, USA

[2] Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL, USA

[3] Department of Psychology, Northeastern University, Boston, MA, USA

[4] Department of Psychology, University of South Carolina Upstate, Spartanburg, SC, USA

[5] Department of Psychology, University of Denver, Denver, CO, USA

[6] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[1] In this paper, we use "attitudes" to refer to latent evaluative knowledge and "evaluations" to refer to observable behaviors, such as self-reports on a Likert scale or binary choices on an AMP. We use "explicit evaluations" to refer to the relatively controlled retrieval of evaluative knowledge and "implicit evaluations" to refer to the relatively automatic retrieval of evaluative knowledge. At the level of measures, we distinguish between "direct measures" and "indirect measures." We do not assume that direct measures capture exclusively controlled processes or that indirect measures capture exclusively automatic processes; however, we believe that given their relatively controlled nature, direct measures are appropriately characterized as indexing explicit evaluations, and given their relatively automatic nature, indirect measures are appropriately characterized as indexing implicit evaluations.

exclusive focus on controlled processes and toward automatic or implicit processes in social evaluation. Implicit evaluations are retrieved under more automatic conditions than their explicit counterparts (Bargh, 1994; De Houwer & Moors, 2010; Moors, 2016; Moors & De Houwer, 2006), such as in the relative absence of intentionality or awareness. That is, when encountering a socially relevant stimulus (such as a person's face, the logo of a sports team, or a social group label), evaluative knowledge related to the corresponding social entity (such as My Spouse–Nice, The Yankees–Bad, or African Americans–Positive[2]) can be activated spontaneously and influence downstream processes of social behavior, ranging from consumer decisions to hiring and promotion and mate selection to criminal sentencing (Bargh, 1989; Devine, 1989; Fazio et al., 1986; Greenwald & Banaji, 1995).

Usually, the automatic activation of evaluative knowledge about social entities is indexed using indirect measures, which bypass asking participants to intentionally report their evaluations of focal stimuli. One of the most popular indirect measures is the affect misattribution procedure (AMP; Payne et al., 2005). Since its introduction in 2005, the AMP has been used to capture implicit evaluations in hundreds of experiments across multiple areas of psychological inquiry, including impression formation (Cone & Ferguson, 2015; Mann & Ferguson, 2015), intergroup relations (Cooley & Payne, 2016; Lee et al., 2018), political behavior (Greenwald et al., 2009; Payne et al., 2010), social cognitive development (Perszyk et al., 2019; Williams & Steele, 2017), and psychopathology (Schreiber et al., 2014; Tucker et al., 2018). Although evidence for the construct validity of the AMP has been accumulating since its introduction (for a review, see Payne & Lundberg, 2014), recent work has challenged the AMP's status as a measure of implicit evaluations.

How does the AMP work? Specifically, what are its methodological features that are seen as critical for its ability to measure implicit evaluations? On each trial of the AMP, two stimuli are presented in rapid succession, first a prime and then a target. The primes are the focal stimuli of interest—which can range from a face to a sports team logo to a social group label—whereas the targets are evaluatively neutral, including stimuli such as Chinese ideographs (for non-Chinese speakers; Payne et al., 2005), inkblots (Williams & Steele, 2017), or abstract paintings (Katz et al., 2022). The critical aspect of this task that makes it indirect is that participants are instructed to rate the pleasantness of the target stimuli and to avoid any biasing influence of the primes.

Despite these instructions, evaluations of the primes are reflected in ratings of the targets, such that targets preceded by positive primes tend to be rated positively, and targets preceded by negative primes tend to be rated negatively. The fact that participants' reactions to the primes influence their reactions to the targets despite instructions to avoid such bias suggests that the evaluations (of the primes) captured by the AMP qualify as unintentional. This methodological feature of the AMP has led to it being used widely as an indirect measure capturing unintentional evaluations, and a substantial body of research supports this claim (Bar-Anan & Nosek, 2016; Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2012; but see Bar-Anan & Nosek, 2011).

At the same time, there has been little empirical work testing the role of other features relevant to the automatic nature of the AMP. In a recent paper, Hughes et al. (2023) examined the role of awareness and argued that, based on these findings, the AMP does not reveal implicit evaluations after all. These authors conclude that the AMP effect depends on "influence awareness,"[3] that is, conscious awareness of the prime's influence on one's response to the target. According to Hughes et al. (2023), this finding suggests that evaluations captured by the AMP are not implicit in the way that has been claimed—a revelation that would upend dozens if not hundreds of published findings using this measure. Given the magnitude of the literature that has been built on the assumption that the AMP measures implicit evaluations, more work is needed to investigate exactly what the AMP does and does not measure. In other words, what methodological features of the AMP are critical in producing responses that can be considered implicit? The aim of the present work is to address this question.

How the AMP measures unintentional evaluations

In what sense are the evaluations captured by the AMP supposed to be implicit? Put another way: What feature of automaticity is supposed to characterize the evaluations measured by the AMP? Some classic theories of automaticity (Shiffrin & Schneider, 1977), dual-process theories of high-level reasoning (Kahneman, 2003), and early approaches to implicit social cognition (Greenwald & Banaji, 1995) did not

[2] Although these examples use an associative notation, for the purposes of this project we are agnostic regarding the representational format of attitudes, e.g., whether they are associative (The Yankees–Bad) or propositional ("The Yankees are terrible").

[3] Throughout the paper, we follow the terminology established by Hughes et al. (2023) and refer to the finding of larger AMP effects on trials in which participants report being influenced by the primes than on trials in which they do not as an "awareness effect." However, in the paper we show empirically that (*a*) such awareness effects can be an outcome, rather than a cause, of AMP effects and (*b*) such awareness effects need not emerge from privileged first-person access but rather can also be subserved by inferential mechanisms.

explicitly reckon with the possibility that different features of automaticity may dissociate from each other. By contrast, other classic and more recent treatments have emphasized the multifaceted nature of automaticity (e.g., Bargh, 1994; Melnikoff & Bargh, 2018; Moors, 2016; Moors & De Houwer, 2006), and specifically the idea that different processing conditions related to automaticity—awareness, intentionality, controllability, efficiency, and speed—need not be aligned with each other.

Notably, in the initial paper introducing the AMP as a new indirect measure, Payne et al. (2005) sought to establish primarily that evaluations indexed by the AMP occur unintentionally. Specifically, the authors provided evidence that AMP effects persist when participants are warned to resist the biasing influence of the primes and even when they are informed of the exact nature of the biasing influence. These conditions should eliminate the AMP effect if the AMP effect depends on the presence of an intention to evaluate the primes (Murphy & Zajonc, 1993). The fact the AMP effect persisted even under these conditions suggests that the AMP captures evaluations that are implicit in the sense of unintentional.

Skepticism about the implicit (unintentional) nature of the AMP effect was voiced by Bar-Anan and Nosek (2011), who demonstrated a statistical relationship between the size of the AMP effect and a post hoc self-report measure of intentional responding to the primes. However, Payne et al. (2012) noted that a mere correlation between the AMP effect and self-reports of intentional responding does not imply that intentional responding plays a causal role in producing the AMP effect. Payne et al. (2012) also presented empirical data challenging the causal inference drawn by Bar-Anan and Nosek (2011).

First, Payne et al. (2012) found that the AMP effect was no more correlated with self-reported intentional responding than with self-reported unintentional responding. Second, expressly encouraging participants to report their intentional evaluations of the primes produced systematic changes in the AMP effect—a finding that is incompatible with the possibility that participants had been evaluating the primes intentionally anyway. In their response, Bar-Anan and Nosek (2016) acknowledged the limitations of the correlational approach to the intentionality of the AMP, and specifically the fact that larger AMP scores may have led to higher self-reports of intentional (and unintentional) responding rather than vice versa.

Following this initial exchange, multiple additional lines of evidence from experimental modifications to the AMP procedure have emerged in support of the unintentional nature of the evaluations measured by the AMP. Specifically, Gawronski and Ye (2013) conducted several studies to rule out two potential alternative mechanisms that may have been responsible for AMP effects: (*a*) prepotent motor responses (by varying the assignment of pleasant and unpleasant responses to different response keys on a trial-by-trial basis) and (*b*) biased perception of the target (by varying the order of primes and targets). In follow-up work, the same authors demonstrated that AMP effects remained intact but were uncorrelated with self-reports of intentional responding when participants (*a*) lacked meta-cognitive knowledge of the source of their evaluative responses because attitudes were induced using mere exposure or (*b*) did not attend to evaluatively relevant features of the primes (Gawronski & Ye, 2015).

However, given that prior work by Payne et al. (2012) showed that reports of (un)intentionality can be confabulated, these findings are not conclusive. Mann et al. (2019) provided additional evidence against intentional evaluation of the primes as a basis of AMP effects. Specifically, these authors probed whether the bimodality occasionally characterizing AMP effect distributions might be the result of intentional responding. Alleviating these concerns, the AMP reflected impression formation effects even on a version of the measure that eliminated such bimodality via the use of alternative instructions, an alternative set of target stimuli, and skip trials (see below).

The role of awareness in the AMP

Given the multifaceted nature of automaticity (Bargh, 1994; Melnikoff & Bargh, 2018; Moors, 2016; Moors & De Houwer, 2006), investigations of intentionality have little bearing on the role of other features of automaticity in the AMP, including awareness—the feature emphasized by Hughes et al. (2023). To use a simple example, someone on a diet may unintentionally experience a positive response as they see a delicious éclair in the window of a bakery. In this situation, the person may or may not be aware of their reaction, but in either case the behavior might be characterized as implicit (automatic) given that it occurs in the absence of any intention on the person's part. It is for this reason that, when introducing the AMP, Payne et al. (2005) left open the question of whether participants are aware of AMP effects, arguing that the issue of awareness "[…] will need to be directly tested in future research" (p. 290).

Although the AMP was originally designed to capture unintentional evaluations, the role of awareness is especially pertinent for the AMP because the measure operates via misattribution. Misattribution of the response from the prime to the target is usually assumed to be possible precisely because people are not aware of that mistake (see Payne et al., 2005). Any evidence to the contrary would be relevant to understanding not only exactly what type(s) of process the measure is indexing, and in particular whether and how the AMP is an indirect measure capturing implicit

evaluations, but also how misattribution might be working in this case.

Because there is already substantial evidence to suggest that the AMP captures largely unintentional evaluations, the potential discovery that the measure might require awareness of the process of misattribution raises the interesting possibility that people might be unintentionally—but consciously—misattributing their reaction to the prime to their evaluation of the target. This effect may be akin to the phenomenon of someone being aware as they yank their hand away from a hot stove. The reaction happens spontaneously and without intention (and perhaps uncontrollably) even while the person is aware of it happening. Based on these considerations, a careful and systematic examination of the role of theoretically relevant methodological features of the AMP procedure is necessary to advance our understanding of the psychological nature of responses on this measure.

Awareness of AMP effects: The recent studies by Hughes et al. (2023)

To summarize, the paper by Payne et al. (2005) that first introduced the AMP highlighted the unintentional nature of the evaluations that it captures as its defining automaticity feature. Accordingly, subsequent investigations of the validity of the AMP as a measure of implicit evaluations have focused on the issue of whether evaluations of the primes occur unintentionally. The overwhelming majority of relevant studies have provided evidence in favor of the implicit (unintentional) nature of the AMP (Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2005, 2012). The sole exception is the study by Bar-Anan and Nosek (2011), whose authors later clarified that they "[…] prefer the illusory intention account [i.e., the possibility that AMP effects lead to self-reports of intentional responding rather than vice versa] as it is more interesting theoretically and preserves the AMP's status as an indirect measure" (Bar-Anan & Nosek, 2016, p. 3). However, this literature so far has left unanswered the question about the possible role of awareness in the AMP effect.

Indeed, given the primary focus on intentionality, to date little research has directly investigated whether the AMP allows for participants to be aware, on a trial-by-trial basis, of the biasing influence of the prime on their response to the target on the AMP. In one rare study, Payne et al. (2012; Experiment 3) found no statistically significant difference between two versions of the AMP, one following the original procedure and one in which participants were allowed to skip trials in which they felt that the prime may have influenced their response. Similarly, Mann et al. (2019) implemented a version of the AMP with the possibility of skipping and found that the impression formation effects observed using the original AMP persisted on this version of the measure. Together, these results have been interpreted to suggest that AMP effects might emerge in the absence of awareness.

However, in their recent investigation of the awareness of AMP effects, Hughes et al. (2023) correctly pointed out several limitations of this approach. Specifically, it is not statistically warranted to conclude that the null hypothesis is true (i.e., that the two versions of the AMP do not differ from each other) from a nonsignificant statistical test (Dienes, 2014; Kruschke, 2018). When Hughes et al. (2023) replicated Experiment 3 of Payne et al. (2012) in a more highly powered within-participant design, they found a statistically significant difference between the two conditions such that AMP effects were larger in the no-skip than in the skip condition. In fact, in the skip condition, the standard AMP effect was reversed.

More importantly, requiring participants to choose between skipping and emitting a response makes it impossible to know how they would have responded on trials that they decided to skip. Driven by this consideration, Hughes et al. (2023) focused on a modified version of the AMP in which, following each trial, the participant is asked to indicate whether their response was influenced by the prime or not. As such, critically, this measure is a self-report measure of awareness, not intention. Even assuming that participants are able to correctly distinguish intentional from unintentional influences (which may well not be the case; see Payne et al., 2012), their task in the Hughes et al. (2023) studies was not to indicate whether they evaluated the prime intentionally. Rather, they were asked to report whether their evaluation of the prime, intentional or not, influenced their subsequent response.

The experiments by Hughes et al. (2023) generated three main sets of empirical findings. (*1*) Most importantly, these authors established an awareness effect such that the AMP effect (i.e., the difference in pleasant vs. unpleasant responding following positive vs. negative primes) was larger for trials in which participants indicated that the prime influenced their responses to the target than for trials in which they did not report such an influence. (*2*) A stable individual difference emerged in rates of influence awareness such that rates of influence awareness in one AMP (using normatively positive and negative images as primes) were correlated with influence awareness rates in a different AMP (using the faces of Barack Obama and Donald Trump as primes). Finally, (*3*) an awareness effect was present even when the measure of awareness was collected (*i*) after exposure to the prime and target but before the participant emitted a response and (*ii*) after exposure to the prime but before the participant saw the target or emitted a response.

This pattern of findings led Hughes et al. (2023) to conclude that the AMP does not measure what it is intended to

measure: "[…] AMP effects are not implicit in the way that has been claimed […]" (p. 1558). In fact, according to these authors, AMP effects are "[…] heavily dependent on influence awareness" (p. 1573), which purportedly suggests that "[…] what is useful about the [AMP] effect is not particularly implicit, and what is implicit is not particularly useful" (p. 1584). In other words, Hughes et al. (2023) argue that the three effects reviewed above call into question the validity of the AMP as a measure of implicit evaluations.

We respectfully disagree with each of these conclusions and believe that the findings of Hughes et al. (2023) provide no evidence that the AMP measures anything other than what it was intended to measure—unintentional evaluations. Even if we were to grant that AMP effects depend on awareness of the prime's influence—which we do not—this claim concerns one automaticity feature that is not perfectly aligned with intentionality. Therefore, any finding of influence awareness would not override the considerable existing evidence suggesting that the AMP measures unintentional evaluations.

At the same time, we agree with Hughes et al. (2023) that identifying the role of awareness in the AMP is important. Especially given recent attention to the feature of unawareness in theoretical work on implicit social cognition (Gawronski et al., 2022), the question of whether AMP effects emerge in the absence of awareness is interesting in its own right. As such, it is worth evaluating the claims that influence awareness both occurs in the AMP and moderates the AMP effect. Compelling evidence for these claims would address a puzzle that, for the most part, has been ignored since the AMP's inception: whether the AMP can be used to draw inferences about evaluations that are in any sense unconscious. Specifically, the present experiments were designed to test the questions of whether (*a*) participants are aware of the prime stimulus as a source of their responses on the AMP and (*b*) whether such awareness (if present) is causally responsible for the AMP effect.

Such additional inquiries are necessary because, even on the issue of awareness, the results of Hughes et al. (2023) raise more questions than they answer: They establish neither an effect of influence awareness on AMP effects, nor even the presence of genuine influence awareness in the first place. Recall that the central finding from Hughes et al. (2023) is a positive correlation between the strength of the AMP effect and self-reported awareness of a prime's influence. From this correlational finding, Hughes et al. (2023) draw a causal inference, for example when stating that "[…] even this purportedly 'improved' version of the task is also heavily dependent on influence awareness" (p. 1573). Plainly, this inference is invalid, given that the data, by virtue of being correlational, are equally consistent with the possibility that awareness has no causal effect on AMP scores at all. For instance, it could be the case that AMP effects are not consequences of self-reported awareness, but rather causes of self-reported awareness (indeed, this is how past work interpreted post hoc self-reports of intentionality; Payne et al., 2012).

Specifically, participants may make post hoc inferences about whether or not they were influenced based on the observed congruency between the primes and their own ratings of the target images: If the two match in valence (i.e., if a pleasant rating of a target image follows a positive prime or an unpleasant rating of a target image follows a negative prime), participants may be likelier to infer that their rating was influenced than if the two mismatch in valence (i.e., an unpleasant rating of a target image follows a positive prime or a pleasant rating of a target image follows a negative prime). This inferential account of Hughes et al.'s findings (briefly acknowledged by the authors themselves in response to reviewer feedback, pp. 1580–1581) supposes that some participants possess the lay theory that a response to the target on an AMP trial is likelier to have been influenced by the prime if the prime and the response are evaluatively congruent with each other.

Importantly, on the inferential account, self-reports of having been influenced need not reflect genuine awareness in the sense of direct, introspective access to the causal effect of the prime on one's own responses—rather, they constitute speculations informed by externally observable events (Bem, 1972; for a similar argument in the context of participants' ability to predict their implicit evaluations, see Morris & Kurdi, 2023). Accordingly, the effects observed by Hughes et al. (2023) are consistent with the notion of unawareness usually used in implicit social cognition research, positing lack of accurate introspective access to (the source of) one's evaluative knowledge (e.g., Greenwald & Banaji, 1995).

The inferential account explains all key findings obtained by Hughes et al. (2023): the fact that judgments of awareness and the size of the AMP effect are statistically related, and the result that participants show stable individual differences in their rates of influence awareness. With regard to the latter, participants who hold a (strong) lay theory about evaluative congruency will show (strong) awareness effects and those who do not will show no awareness effects. Under this account, participants' awareness rates on one AMP are predictive of their AMP effect on a different AMP not because awareness rates cause the AMP effect on both. Rather, participants exhibit stable individual differences in their beliefs about evaluative congruency, and therefore, awareness rates across different AMPs will be correlated with each other.

Two findings from Hughes et al. (2023) may, at a first glance, appear inconsistent with the inferential account, but are not. First, Hughes and colleagues found that the awareness effect emerges even if awareness is measured before participants rate the target image. This finding may seem to rule out the possibility that awareness reports are based

on the match or mismatch between target ratings and prime valence. This is not the case. As recognized by Hughes et al. (2023) themselves, participants may internally evaluate the targets before pressing the corresponding key on their keyboard (indeed, this is precisely what is supposed to happen on the view that the primes are evaluated unintentionally) and, based on these internal evaluations, infer whether they were influenced by the primes. Accordingly, measuring awareness prior to the collection of target ratings provides no evidence about the causal direction of the effect between awareness and the AMP effect.

Second, Hughes and colleagues found that the AMP effect is larger on aware than on unaware trials even when awareness reports are collected before the target is even shown. This finding can be explained straightforwardly in terms of consistency motives (Festinger & Carlsmith, 1959; Gawronski, 2012) to strategically keep target ratings evaluatively congruent with the prime conditional on having indicated influence awareness. However, this type of self-fulfilling prophecy is likely to be specific to the design in which the influence judgment precedes presentation of the target and thus not particularly informative with respect to the operating conditions of the original measure. After all, neither the standard AMP nor any of the modified procedures implemented by Hughes et al. (2023) involve soliciting a precommitment about their response from participants.

To summarize, Hughes et al. (2023) found larger AMP effects when participants reported that they were influence-aware than when they did not, and from this finding, concluded that the AMP is not a valid measure of implicit evaluations. However, as we have argued, the characterization of the AMP as an indirect measure has depended primarily on its ability to capture unintentional evaluations, without any commitment to the role of awareness. As such, we agree with Hughes et al. (2023) that it is important to define, in each individual case, whether the label "implicit" is used to refer to unintentional or unaware evaluations (or automatic evaluations in some other sense; see the hot stove example above). However, the fact that participants may be aware of AMP effects need not negate the status of the AMP as a measure of implicit evaluations. Moreover, even putting this theoretical concern aside, the correlational findings reported by Hughes and colleagues leave open reverse causality and third-variable explanations, which we examine in the present research.

The present project

Recent work by Hughes et al. (2023) has been interpreted to suggest that one of the most widely used indirect measures in social cognition—the AMP—is, in fact, producing responses that are misaligned with its status as an indirect measure. The present project offers an analysis of these claims by systematically varying methodological features of the AMP under theoretically relevant conditions.

Our response to Hughes et al.'s (2023) claims about the AMP as a measure of implicit evaluations boils down to two arguments, one empirical and one conceptual. The conceptual argument is that the results of Hughes et al. (2023) do not, and cannot in principle, have any bearing on the AMP's status as a measure of unintentional evaluations given that Hughes et al. (2023) did not investigate this feature of automaticity. The studies by Hughes et al. (2023) may, however, be seen as relevant to another (not perfectly aligned) feature of automaticity—awareness. Although awareness has been treated as secondary in importance to intentionality in past work on the AMP, it can still provide valuable information on the way(s) in which the AMP is or is not automatic.

More importantly, our empirical argument is that the awareness effect documented by Hughes et al. (2023) via modifications of the AMP procedure can be explained in terms of an inferential account, according to which genuine influence awareness neither moderates the AMP effect, nor even occurs in the first place. We provide evidence for these ideas across two sets of experiments: In one set (Experiments 1–4) we modify methodological features of the AMP to test whether the awareness effect documented by Hughes et al. (2023) is replicable and generalizable across variations in task parameters, and in the second set (Experiments 5–7) we introduce experimental manipulations to the AMP to directly test the inferential account. Together, these studies both address the claims from Hughes et al. (2023) about the role of awareness in the AMP and, more generally, provide a systematic analysis of the role of multiple elements of the AMP procedure in producing (automatic) evaluative responses. In doing so, this work presents new evidence about the importance of various procedural features of the AMP, which is necessary for advancing the field's understanding of the processes that this measure does and does not capture.

In Experiment 1, we conducted a close replication of Experiment 2 from Hughes et al. (2023), using the modified AMP that relies on highly valenced (extremely positive and extremely negative) primes and includes an influence awareness measure following each trial. Specifically, similar to the original study, participants were asked to respond if they thought that their response to the target was influenced by the prime (go response) and to withhold a response if it was not (no-go response). The goal of this experiment was to establish the replicability of the awareness effect, i.e., a larger AMP effect on trials in which participants reported being influenced by the primes than on trials in which they did not.

In the remaining experiments in the first set, we turned to investigating the robustness and boundary conditions of the

awareness effect by varying different elements of the modified AMP procedure. Together, these experiments provide direct evidence on the scope of the applicability of Hughes et al.'s (2023) claims and the size of the awareness effect, along with suggestive indirect evidence on the cognitive mechanism(s) giving rise to self-reports of influence awareness. In Experiment 2, we reversed the contingencies for go versus no-go responses relative to Experiment 1. That is, participants were asked to respond if they were not influenced by the prime and to withhold a response if they were influenced. In Experiment 3, we removed the asymmetry between go and no-go responses by asking participants to choose between two different responses to indicate influence versus lack thereof.

If, as suggested by Hughes et al. (2023), self-reports of prime influence are to be taken as direct indications of participants' privileged first-person awareness of their mental processes, these procedural variations in response format should not produce major differences in self-reported rates of awareness. However, if self-reports of prime influence emerge from informationally promiscuous inferential processes that reflect myriad sources of evidence (Bem, 1972; Morris & Kurdi, 2023; Payne et al., 2012), then response format may be expected to modulate influence awareness in the direction suggested by the default option.

Lastly, in Experiment 4, we varied prime valence continuously from highly negative to highly positive rather than relying on only extremely valenced primes, as Hughes et al. (2023) did. This experiment is important because the vast majority of the literature using the AMP uses primes that are relatively mild in valence, such as human faces (Cone & Ferguson, 2015; Mann & Ferguson, 2015), groups of people (Cooley & Payne, 2016), or words or images related to mental illness (Schreiber et al., 2014; Tucker et al., 2018). Therefore, it is important to probe whether self-reports of awareness—even if emerging from an inferential mechanism—are as prevalent for non-extreme as for extreme primes. After all, if the awareness effect reported by Hughes et al. (2023) is limited to evaluatively extreme primes, then this would diminish the seriousness of any purported threat to the validity of the AMP raised by those authors to a considerable degree even if their casual account is accurate.

In the second set of experiments, we more directly probed the mental processes underlying self-reports of prime influence, and more specifically the possibility that the AMP effect may cause influence awareness rather than vice versa (the "inferential account"). In Experiment 5, we asked participants to judge influence awareness on a trial-by-trial basis not only for themselves but also, on a separate task, for a participant from a previous study. In this experiment, we probed whether (*a*) the awareness effect emerges on the third-party AMP and (*b*) awareness effects on the first-person and third-party AMPs are related to each other. In Experiment 6, participants made judgments of prime influence on a modified version of the AMP in which primes were presented only post hoc, not during the actual trials. As such, in Experiments 5–6, an awareness effect would be expected to emerge if influence judgments are caused by AMP effects (as predicted by the inferential account) but not if, as proposed by Hughes et al. (2023), AMP effects are caused by influence awareness.

Finally, according to Hughes et al. (2023), a misattribution mechanism—which is widely assumed to give rise to AMP effects—is difficult to reconcile with participants being aware of the influence of the primes on their responding to the targets. In Experiment 7, we directly evaluated this claim. Specifically, we manipulated the evaluative strength of the primes (neutral, mildly negative, and extremely negative), and asked participants to report (*a*) to what extent their response was influenced by the prime (as they had done in the studies by Hughes and colleagues) and (*b*), newly, to what extent their response was influenced by the target. A parallel effect of prime strength on judgments of prime influence and target influence follows from the inferential account, but not from an account on which awareness of priming causes the AMP effect.

Experiment 1

Relying on data from seven experiments, Hughes et al. (2023) reported an awareness effect on the affect misattribution procedure (AMP; Payne et al., 2005): AMP effects were larger on trials where participants indicated that their response to the target had been influenced by the prime than on trials where they did not. Given the novelty of this procedure and the findings emerging from it, in Experiment 1 we conducted a close replication to examine their robustness. Although the results are inconclusive with respect to the underlying theoretical mechanism, this experiment constitutes a first step toward establishing the replicability of the awareness effect—a prerequisite for probing its generalizability (Experiments 2–4) and the cognitive processes giving rise to it (Experiments 5–7).

Method

Open science practices

For all experiments, we report how we determined the sample size as well as all data exclusions, manipulations, and measures. In Experiments 1–6, sample size was determined based on the common practice in the senior author's lab of recruiting at least 150 participants per between-participant condition, with additional participants recruited to account for participant exclusions. A priori power analyses were

conducted for Experiment 7. Specifically, the hypotheses in this experiment concerned the within-participants effect of prime extremity on judgments of influence, separately in the prime-influence and target-influence conditions. Accordingly, we estimated that a sample size of 150 in each condition would provide adequate power (0.80) to detect a small within-participants effect of prime extremity ($d_z = 0.15$) within each condition. No intermittent data analyses were conducted in any experiment.

Sample sizes, exclusion criteria, designs, and analysis plans were preregistered for Experiments 2–7. Due to a clerical error, in Experiments 2–6 we did not preregister that we would exclude participants with zero variance in AMP responses. Given that such participants are routinely excluded from analyses in work relying on the AMP, we excluded these participants from the main analyses reported in the paper. However, as shown in the open code, none of the inferential results would change with these participants retained. Any further deviations from the preregistration are noted in the relevant Method sections.

All materials, stimuli, data (including trial-level AMP data), and analysis scripts are available for download from the Open Science Framework (OSF; https://osf.io/wfksp/).

Participants and design

A total of 375 volunteer participants were recruited from the Project Implicit educational website (https://implicit.harvard.edu/implicit/). Participants were excluded from analyses if they did not complete the affect misattribution procedure (AMP; $n = 68$) or pressed the same key on all AMP trials, indicating noncompliance with instructions ($n = 34$). In line with Hughes et al. (2023), we also collected a self-report measure identifying participants who believed that their data were unusable. The results were highly similar with these participants removed or retained; as such, we retained these participants for analysis. Given that 12 participants were excluded based on more than one of the criteria described above (e.g., they had both an incomplete AMP and pressed the same key on all AMP trials), the final sample size was 285.[4]

Participants represented a total of 31 countries of citizenship, with the United States accounting for most of the sample ($n = 193$, or 68%) and no other country accounting for more than 6%. Eighty-three percent of participants were from majority-English-speaking countries, including the United States, Canada, the United Kingdom, Australia, South Africa, New Zealand, and Ireland.[5] In the final sample, 187 participants were women, 90 were men, and 5 of other genders. The mean age was 39 years ($SD = 14$).

All participants completed the modified AMP introduced in Experiment 2 of Hughes et al. (2023), in which they indicated, following each trial, whether their response to the target was influenced by the prime. Prime valence (positive vs. negative) was manipulated within participants, and influence awareness (unaware vs. aware) was measured on each trial.

Materials

Valenced images The same set of valenced images used by Hughes et al. (2023) and originally obtained from the International Affective Picture System (IAPS; Lang et al., 2008) were retained for use as primes on the AMP. The set included 12 highly positive and 12 highly negative images.[6]

Abstract paintings Unlike Hughes et al. (2023), who used Chinese ideographs, we used a set of 80 abstract paintings created by Katz et al. (2022) as targets on the AMP. In line with the logic of the AMP procedure, these images were evaluatively ambiguous (see Mann et al., 2019; Study 3A). Most recent work in the senior author's lab has relied on abstract paintings rather than Chinese ideographs as target stimuli because this procedure does not require excluding participants based on language proficiency.

Procedure and measures

Affect misattribution procedure The initial instructions for the affect misattribution procedure (AMP; Payne et al., 2005) were modeled after those used by Hughes et al. (2023). Participants were informed that during the task, they would see pairs of images appear one after the other, specifically that a real-life image (prime) would be followed by an abstract painting (target). They were asked to judge how pleasant each target was while ignoring the primes. Participants were told that the primes might sometimes bias their judgments of the targets and were asked to do their absolute best in trying to resist this biasing influence. Furthermore, participants were asked to indicate, following each trial,

[4] A precise breakdown of reasons for participant exclusions, including in cases where multiple criteria led to an exclusion decision, is available in the open data.

[5] An anonymous reviewer expressed concerns about the English proficiency of participants from non-English-speaking countries. To alleviate these concerns, we refit the main models from Experiments 1–5 to the data of participants from majority-English-speaking countries only and found no substantial deviation from the conclusions reported in the paper. (Participants in Experiments 6–7 were recruited exclusively from the United States.) The corresponding models are available in the open code.

[6] IAPS images were not originally intended for use in online research (Lang et al., 2008). However, given that these images were used as primes in the online experiments conducted by Hughes et al. (2023), we retained them for use in the present studies.

whether the prime influenced their judgment of the target by pressing the space bar if they were influenced and by waiting for the next trial to start if they were not.

The AMP consisted of 80 trials. Each trial included the presentation of (*a*) a fixation cross for 500 ms, (*b*) a prime (valenced image) for 100 ms, (*c*) a blank screen for 100 ms, (*d*) a target (abstract painting) for 100 ms, and (*e*) a black-and-white pattern mask until the participant entered a response. Participants used the "E" key on their keyboard to indicate that the target was less pleasant than average and the "I" key to indicate that the target was more pleasant than average. Following 10 practice trials that were removed from analyses, positive IAPS images served as primes on 35 trials (positive primes condition) and negative IAPS images served as primes on 35 trials (negative primes condition). Primes were sampled randomly, without replacement. Once the full set of primes was exhausted, sampling without replacement began anew. Each target stimulus was used on a single trial of the AMP. The order of AMP trials was individually randomized.

Following each trial, participants made a judgment of whether the prime influenced their evaluation of the target. If they thought it did, they were asked to press the space bar. If not, they were asked to wait for the next trial to start. The influence judgment screen was displayed for a fixed duration of 2300 ms, irrespective of whether the participant pressed the space bar or not. However, in line with Hughes et al. (2023), participants were not informed of this fact.

Overall, the AMP procedure emulated the procedure of the modified AMP from Experiment 2 of Hughes et al. (2023), with the exception that, in line with the change in target stimuli, the instructions referred to abstract paintings rather than to Chinese symbols. In addition, the procedure of the experimental script shared by Hughes et al. (2023) on OSF deviated from the Method section of their Experiment 2 in two ways. Specifically, (*a*) as revealed by the open data posted by Hughes et al. (2023) on OSF (https://osf.io/8mrny/), instead of 10 practice trials followed by 120 trials, the AMP consisted of a total of 80 trials (8 practice trials and 72 experimental trials, which we modified to 10 practice trials and 70 experimental trials in the present work), and (*b*) the response deadline for the influence judgments was fixed to a duration of 2300 ms rather than 2000 ms. In these cases, we followed the experimental script rather than the Method section of Hughes et al. (2023).

Exploratory measures Following the AMP, participants completed a set of explicit items, asking them to report (*a*) the proportion of trials in which a prime was present; (*b*) the proportion of trials in which a target was present; (*c*) the proportion of trials in which their response to the target was influenced by the prime; (*d*) the proportion of trials in which their response to the target was unintentionally influenced by the prime; and (*e*) the proportion of trials in which their response to the target was intentionally influenced by the prime; to describe, in a free response format, (*f*) the strategy that they used to decide whether their response to the target was influenced by the prime, and (*g*) the strategy that they used on trials in which they were unaware of the prime or the target; as well as to indicate (*h*) whether their data should or should not be used in statistical analyses. These items were collected for exploratory purposes and are not discussed further; however, they are available in the open data.

Analytic strategy

Statistical analyses for this and all remaining experiments were performed in the R statistical computing environment (version 4.0.3). Unless otherwise noted, inferential analyses were conducted at the trial level via linear mixed-effects modeling using the lme4 package (version 1.1.23; Bates et al., 2015). In Experiments 1–4, the binary response variable (1 = pleasant; 0 = unpleasant),[7] in Experiments 5–6 the binary awareness variable (1 = aware; 0 = unaware), and in Experiment 7 the continuous awareness variable served as the dependent variable.

The reason for the change in dependent variables was a shift in focus: Experiments 1–4 investigated the replicability and generalizability of Hughes et al.'s (2023) original findings, who treated the binary AMP response as the dependent variable; by contrast, Experiments 5–7 newly probed the cognitive mechanisms from which self-reports of influence awareness emerge, which made it reasonable to treat this variable as the dependent measure. A logit link function was used to model the binary dependent variables in Experiments 1–6, whereas in Experiment 7, no link function was used given that the dependent variable was numeric.

Model fitting proceeded stepwise. Specifically, the random part of the model was added first and the fixed part of the model second, with likelihood ratio tests calculated after each modeling step to ascertain incremental gains in model fit.[8] For reasons of parsimony and computational viability,

[7] For Experiments 2–4, due to a clerical error, the preregistration documents list the awareness variable, rather than the AMP response, as the dependent variable. We follow the procedure of Hughes et al. (2023) in treating the awareness variable, rather than the AMP response, as the dependent variable. However, the substantive conclusions would remain unchanged even if the AMP response were treated as the dependent variable.

[8] Given the stepwise model comparison process, the degrees of freedom reported for the likelihood ratio test can differ from experiment to experiment even if the best-fitting model is the same. The reason for this is that the likelihood ratio test takes into account not only the complexity (number of parameters) of the final model but also the complexity (number of parameters) of the penultimate model. For example, the degrees of freedom for the likelihood ratio test will be larger when the penultimate model contains only one main effect rather than two main effects. We thank an anonymous reviewer for bringing this point to our attention.

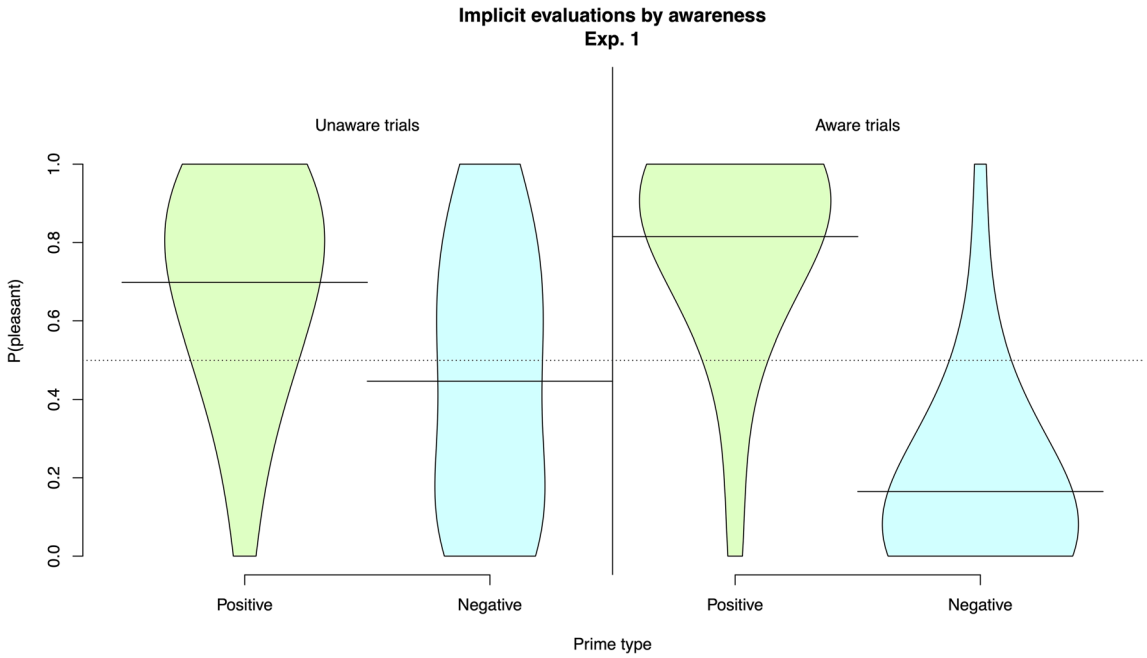

Fig. 1 Implicit (AMP) evaluations by prime type (positive vs. negative) and self-reported influence awareness at the trial level (unaware vs. aware; Experiment 1). The *y*-axis shows the proportion of pleasant responses, with the dashed line at 0.5 marking neutrality. The solid lines represent condition means

only random intercepts and no random slopes were used.[9] The best-fitting model was retained and interpreted. For the sake of brevity, not all model fitting steps are reported in the paper, but all are available in the open code. Planned comparisons were conducted using the emmeans package (version 1.5.0; Lenth, 2020).

Results

The proportion of pleasant responses on the AMP by prime condition (positive vs. negative) and self-reported influence awareness (unaware vs. aware) is shown in Fig. 1.

Participants were more likely to rate the target as pleasant following positive, $P(\text{pleasant}) = 0.732$, than negative primes, $P(\text{pleasant}) = 0.356$, resulting in a mean AMP effect of $P_{\text{diff}}(\text{pleasant}) = 0.375$ (median = 0.371; $SD = 0.373$). The mean proportion of aware trials at the participant level was $P(\text{aware}) = 0.316$ (median = 0.171; $SD = 0.342$). Importantly, in line with Hughes et al. (2023), participants showed a larger AMP effect on aware, $P_{\text{diff}}(\text{pleasant}) = 0.667$, than on unaware trials, $P_{\text{diff}}(\text{pleasant}) = 0.238$. Notably, as indicated by the latter effect size, the AMP effect was robust even on unaware trials. Moreover, given that the number of unaware trials exceeded the number of aware trials by an approximately 2:1 ratio, the overall AMP effect was numerically considerably closer to the unaware than to the aware AMP effect.

Accordingly, the best-fitting model with valence of response (pleasant vs. unpleasant) as the dependent variable contained random intercepts for participants, primes, and targets, as well as a Prime Type (positive vs.

[9] In response to reviewer feedback, we conducted exploratory analyses to investigate whether including the maximal participant-level random-effects structure in each model leads to similar conclusions as the more parsimonious models reported in the main text.

To test for model overparameterization, we followed the procedure recommended by Bates et al. (2018) and Matuschek et al. (2017) and conducted a principal component analysis on the random-effects covariance matrix of the full model, and then simplified the model if there were principal components accounting for 0% of the variance after rounding to three decimal places. We simplified the model by first disallowing correlation between the random effects; then, if there was still a degenerate component, we dropped the corresponding random slope.

In Experiment 4, we were unable to follow this approach given that the main model was a generalized additive mixed-effects model. In Experiments 1–3 and 7, the statistical inferences remained unchanged. In Experiments 5 and 6, the best-fitting models remained the same, with minor deviations in planned comparisons. Specifically, unlike in the main models, the first-person awareness effect with respect to positive stimuli was reduced to non-significance in the maximal model of the Experiment 5 data, and the awareness effect was reduced to non-significance with respect to positive stimuli in the maximal model of the Experiment 6 data. Both of these results are indicative of a valence asymmetry effect observed both in the original experiments of Hughes et al. (2023) and in the present work. We return to this effect in the general discussion.

negative) × Awareness (unaware vs. aware) interaction as the fixed effect, $\chi^2(1) = 867.87$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that the strength of the AMP effect differed significantly across the two levels of the awareness variable. Notably, in planned comparisons, we found a significant AMP effect on both unaware, $b = 1.29$, $z = 8.15$, $p < 0.001$, and aware trials, $b = 3.67$, $z = 22.31$, $p < 0.001$.

Discussion

We replicated the awareness effect established by Hughes et al. (2023) such that, in a modified AMP procedure involving self-reports of prime influence following each trial, the AMP effect was larger on influence-aware than on influence-unaware trials. We also replicated two additional results also obtained but not emphasized by these authors: (*a*) an AMP effect was clearly present even on influence-unaware trials, and (*b*) given that most trials were influence-unaware trials, the overall AMP effect was considerably closer in size to the AMP effect obtained on unaware than on aware trials.

Overall, the results of Experiment 1 suggest that, relative to the size of the overall AMP effect, the awareness effect may be not as large or consequential as previously assumed. However, similar to the experiments conducted by Hughes and colleagues, these results cannot arbitrate between two plausible causal models of these effects: (*a*) the AMP effect being caused by influence awareness (the possibility favored by Hughes et al., 2023) versus (*b*) influence awareness being caused by the AMP effect (the inferential account described in the introduction). Before we turn to direct tests of this crucial theoretical question in the second part of the paper, in Experiments 2–4 we examine the robustness of the awareness effect across different versions of the modified AMP procedure. These experiments provide both indications of generalizability and initial evidence on the cognitive processes that may produce self-reports of influence awareness.

Experiment 2

The procedure used by Hughes et al. (2023) to probe influence awareness had a peculiar feature: participants were asked to press a button if they thought their response was influenced by the prime (go response) and to withhold a response if they thought that it was not (no-go response). Given well-documented default effects in the judgment and decision-making literature (Jachimowicz et al., 2019), i.e., the tendency to choose whatever option has been preselected, including in consequential contexts where decisions have life-and-death consequences (Johnson & Goldstein, 2003), we suspected that the asymmetry between go and no-go responses may have contributed to the results obtained by Hughes et al. (2023) and, by extension, in Experiment 1 above.

As such, Experiment 2 included two conditions: a standard condition mirroring the asymmetry between go and no-go responses from Experiment 1 and a reversed condition in which influence awareness was the default (no-go) option and influence unawareness was the non-default (go) option. We sought to examine whether this manipulation would (*a*) produce an effect on the base rate of aware responses and (*b*) modulate the relationship between self-reports of influence awareness and the AMP effect. As explained in more detail below, the reversed condition was expected to produce a lower percentage of aware responses; we were agnostic as to whether the relationship between influence awareness and the AMP effect would change and, if so, how.

This experiment is theoretically informative in two ways. First, it provides some indication of how robust and reliable the measure of influence awareness used by Hughes et al. (2023) is. Indeed, invalid or unreliable measures can lead to erroneous conclusions about psychological phenomena (Flake & Fried, 2020; Hussey & Hughes, 2019) and, specifically, about awareness of (the source of) one's attitudes (Kurdi et al., 2022a). In this particular case, the default response may have led Hughes et al. (2023) to underestimate the extent of influence awareness on the AMP.

Second, this experiment can provide initial evidence on whether self-reports of prime influence can reasonably be taken to be a direct and unadulterated index of participants' privileged introspective access to their own cognitive processes, as assumed by Hughes et al. (2023). On this view, the default response should exert little if any effect on reports of influence awareness. However, if reports of awareness reflect inferentially promiscuous propositional processes, then they may well incorporate different sources of information, including the default response format. Specifically, under the inferential account, participants may reason that whatever response was the default was also the one expected by the experimenter, thus giving rise to an acquiescence bias (Tourangeau et al., 2000). Similarly, the mere goal of looking for influence awareness may prompt participants to overidentify or misidentify external or internal clues as indicating prime influence (Jones & Sugden, 2001).

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiment 1.

Participants and design

A total of 710 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/implicit/). Participants who started or completed Experiment 1 were not allowed to participate in Experiment 2; to avoid repeat participants, a similar exclusion rule was applied across all remaining experiments relying on the Project Implicit participant pool (Experiments 3–5). Participants were excluded from analyses if they did not complete the AMP ($n = 110$) or pressed the same key on all AMP trials ($n = 78$). Given that 19 participants were excluded based on multiple criteria, the final sample size was 541.

Participants represented a total of 43 countries of citizenship, with the United States accounting for most of the sample ($n = 404$, or 75%) and no other country accounting for more than 5%. Eighty-four percent of participants were from majority-English-speaking countries. In the final sample, 345 participants were women, 178 men, and 9 of other genders. The mean age was 38 years ($SD = 14$).

The design was similar to Experiment 1, with one additional between-participant manipulation. Specifically, participants assigned to the standard condition ($n = 260$) completed the same AMP as in Experiment 1, whereas participants assigned to the reversed condition ($n = 281$) completed a version of that AMP in which they were asked to respond if they thought their response was not influenced by the prime and to withhold a response if they thought there was an influence. In other words, in the reversed condition, the go and no-go responses were flipped. Additionally, as in Experiment 1, prime valence (positive vs. negative) was manipulated within participants, and influence awareness (unaware vs. aware) was measured on each trial.

Procedure and measures

AMP The AMP procedure was identical to the one used in Experiment 1, with the exception that participants in the reversed condition were asked to withhold a response when they thought their response to the target had been influenced by the prime and to press the space bar when they thought their response had not been influenced by the prime.

Exploratory measures In addition to the set of measures collected in Experiment 1, two new exploratory items were added. These items (*a*) asked participants to report the proportion of trials in which they hit the space bar to advance more quickly to the next trial and (*b*) provided participants with a free-response option to provide feedback on any aspect of the experiment. We included the former item because, although the response deadline was fixed to a duration of 2300 ms, participants were not informed of this fact. As such, we anticipated that some participants might assume that they would be able to proceed more quickly through the experiment by indicating awareness (or lack of awareness, depending on condition), thus adding error variance to this crucial measure.

Results

The proportion of pleasant responses by AMP condition (standard vs. reversed), prime condition (positive vs. negative), and self-reported influence awareness (unaware vs. aware) is shown in Fig. 2.

In the standard condition, we replicated the results of Experiment 1 and those of Experiment 2 from Hughes et al. (2023). Specifically, participants were more likely to rate the target as pleasant following positive, $P(\text{pleasant}) = 0.745$, than negative primes, $P(\text{pleasant}) = 0.352$, resulting in a mean AMP effect of $P_{\text{diff}}(\text{pleasant}) = 0.393$ (median = 0.400; $SD = 0.350$). The mean proportion of aware trials at the participant level was $P(\text{aware}) = 0.358$ (median = 0.257; $SD = 0.336$). Importantly, participants showed a larger AMP effect on aware, $P_{\text{diff}}(\text{pleasant}) = 0.662$, than on unaware trials, $P_{\text{diff}}(\text{pleasant}) = 0.239$. As in Experiment 1, an AMP effect was clearly present on unaware trials and the size of the overall AMP effect was similar to the AMP effect on unaware trials.

In the reversed condition, we sought to probe to what degree these results would be malleable in response to reversing the contingencies for go versus no-go responses. Similar to the standard condition, participants were more likely to rate the target as pleasant following positive, $P(\text{pleasant}) = 0.759$, than negative primes, $P(\text{pleasant}) = 0.367$. As such, unsurprisingly, the size of the overall AMP effect was virtually identical to that observed in the standard condition, $P_{\text{diff}}(\text{pleasant}) = 0.392$ (median = 0.429; $SD = 0.372$). However, the mean proportion of aware trials at the participant level increased considerably, by 58% to $P(\text{aware}) = 0.566$ (median = 0.629; $SD = 0.376$), suggesting that the default (no-go) response had a sizable effect on influence judgments. The difference in awareness rates between the conditions was statistically significant and medium in size, $t(538.4) = 6.80$, $p < 0.001$, Cohen's $d = 0.58$.

At the same time, the difference in AMP effects between aware and unaware trials was similar to the standard condition, with a larger AMP effect emerging on aware, $P_{\text{diff}}(\text{pleasant}) = 0.594$, than on unaware trials, $P_{\text{diff}}(\text{pleasant}) = 0.129$. As such, although an AMP effect was clearly present on unaware trials, in this condition, due to the shift in base rates, the size of the overall AMP effect was more similar to the AMP effect on aware trials than to the AMP effect on unaware trials.

Accordingly, the best-fitting model with valence of response (pleasant vs. unpleasant) as the dependent

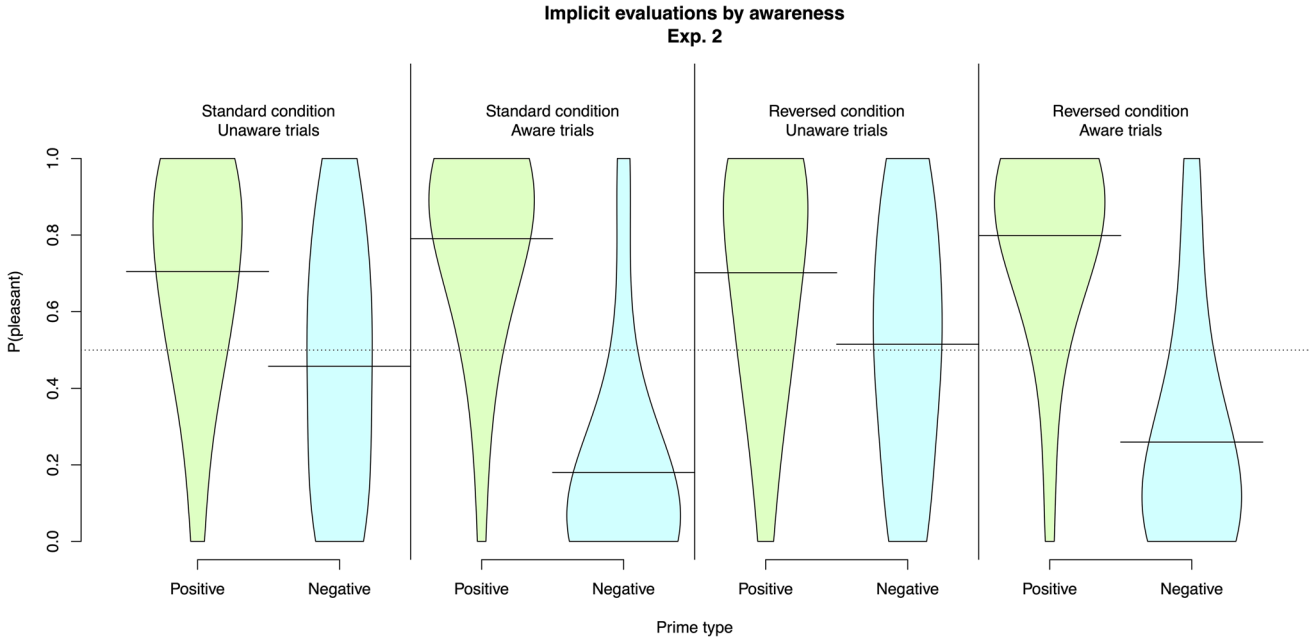

Fig. 2 Implicit (AMP) evaluations by condition (standard vs. reversed), prime type (positive vs. negative), and self-reported influence awareness at the trial level (unaware vs. aware; Experiment 2). The *y*-axis shows the proportion of pleasant responses, with the dashed line at 0.5 marking neutrality. The solid lines represent condition means

variable contained random intercepts for participants, primes, and targets, as well as a Prime Type (positive vs. negative) × Awareness (unaware vs. aware) × AMP Type (standard vs. reversed) interaction as the fixed effect, $\chi^2(3) = 103.88$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that the strength of the AMP effect differed significantly across the two levels of the awareness variable, and that the size of the awareness effect differed across the standard and reversed AMP conditions. Notably, in planned comparisons, we found a significant AMP effect on both unaware, $b = 1.25$, $z = 8.79$, $p < 0.001$, and aware trials, $b = 3.70$, $z = 23.88$, $p < 0.001$, in the standard AMP condition, and on both unaware, $b = 0.73$, $z = 5.07$, $p < 0.001$, and aware trials, $b = 3.20$, $z = 21.92$, $p < 0.001$, in the reversed AMP condition.

In an exploratory analysis, we probed to what extent participants reported pressing the space bar to advance more quickly to the next trial, thus adding error variance to the measure of awareness. The mean of this variable was 26.20 ($SD = 33.89$) on a 100-point scale, indicating considerable levels of contamination, which differed significantly from zero, $t(514) = 17.55$, $p < 0.001$, Cohen's $d = 0.77$. Moreover, given the socially sensitive nature of not following experimenter instructions, we suspect that participants may have underreported, rather than overreported, the extent of this behavior.

Remarkably, self-reported propensity of pressing the space bar to advance more quickly was highly correlated with the proportion of aware responses at the participant level, $r = 0.477$, $t(247) = 8.52$, $p < 0.001$, in the standard condition, and moderately correlated with the proportion of aware responses, $r = -0.217$, $t(264) = -3.61$, $p < 0.001$, in the reversed condition.[10] We believe that these correlations may have been even higher if the measure of skipping behavior had been obtained online (i.e., following each trial) rather than post hoc and if socially desirable responding had been accounted for.

Discussion

The awareness effect emerged both in the standard condition that followed the procedure of Hughes et al. (2023) and Experiment 1 and in the reversed condition, which required a no-go response to indicate influence awareness and a go response to indicate lack of influence awareness. This result suggests that the correlation between AMP effects and reported awareness is robust across manipulations of the default option. However, remarkably, the base rates of aware responses differed considerably across the two conditions (36% vs. 58%); therefore, the overall AMP effect closely

[10] The sign flip between the two correlations is theoretically expected given that in the standard condition, pressing the space bar indicated influence awareness, and in the reversed condition, it indicated lack of influence awareness.

resembled the AMP effect obtained on unaware trials in the standard condition and the AMP effect obtained on aware trials in the reversed condition. This is an important result because, even assuming the accuracy of Hughes et al.'s (2023) causal account—which we do not—it indicates large amounts of uncertainty over the extent to which influence awareness is implicated in AMP effects.

This finding points to serious psychometric issues with the measure of influence awareness, at least if this measure is assumed to be an accurate indication of participants' privileged access to their own cognitive processes: Although the two conditions were logically equivalent to each other and differed only in the default option, participants were 58% more likely in the reversed than in the standard condition to indicate that their response to the target had been influenced by the prime. This amount of variability is inconsistent with the idea that the influence awareness measure provides a pure and stable reflection of privileged introspective processes. Rather, this finding is reminiscent of the results of Experiment 1 from Payne et al. (2012), where participants agreed either with the suggestion that the influence of the primes on their responses was intentional or that it was unintentional, depending on which possibility was mentioned in the question.[11]

We find these results not particularly surprising: The situation created on each trial of the AMP is intentionally ambiguous such that participants can never conclusively determine whether their response was influenced by the prime or not. As mentioned in the introduction, we believe that at least some participants have the lay theory that prime influence is more likely if the prime and the response are evaluatively congruent with each other. However, this theory is, in and of itself, insufficient. The information to which participants would have to be privy, but are not, is how they would have responded to the target had the prime not been present. This is an extremely challenging question, and therefore participants seem to rely on any information available to them, including the default response option and potentially their guesses about the experimenter's expectations, to answer it. This result provides initial evidence for the idea—explored in more detail in the remaining experiments—that self-reports of prime influence emerge from informationally promiscuous inferential processes.

In an exploratory analysis, we identified an additional potential issue with the measure of influence awareness: Although the duration of the trial was the same whether participants pressed the space bar or not, participants had not been informed of this fact. As such, participants reported that about 26% of the time, they had pressed the space bar to be able to advance more quickly to the next trial. Importantly, the self-reported tendency to do so was highly correlated with the proportion of aware responses in the standard condition and moderately correlated with it in the reversed condition. We suspect that, given that this behavior was a clear violation of experimental instructions, the true extent of the issue may be considerably greater. Of course, this correlational finding does not establish the nature of the causal relationship between post hoc self-reports of skipping and online self-reports of influence awareness. However, at the very least, it indicates that participants are willing to accept a wide variety of explanations for their own behavior offered to them by the experimenter. This finding, in turn, is difficult to reconcile with the possibility that judgments of influence awareness emerge from accurate introspection.

Experiment 3

Given (*a*) the large differences in base rates of self-reported influence awareness across the standard and reversed versions of the influence awareness item and (*b*) the correlation between self-reported skipping and the proportion of aware responses at the participant level, in Experiment 3 we removed the asymmetry between go and no-go responses that characterized the experiments by Hughes et al. (2023) and, by extension, Experiments 1–2 above. Specifically, in Experiment 3, following each AMP trial, participants were asked to press one key if they thought that their response had been influenced by the prime and a different key if they thought that it had not. Similar to Experiments 1–2, this experiment sought to determine (*a*) the base rate of aware responses and (*b*) the extent of the relationship between self-reported awareness and the AMP effect.

This experiment served multiple goals. First, given the large default effect observed in Experiment 2, we sought to establish a paradigm to use in the remaining studies that was not subject to this concern. Second, and critically, along with Experiment 2, this study provides additional evidence regarding (*a*) the size of the awareness effect and

[11] An additional explanation of the substantial effect of response options relies on the idea of response sets, that is, responding to self-report items in a way that is determined by the structure rather than the content of the question. Specifically, work involving the evaluative priming procedure suggests that participants are more likely to respond "yes" than "no" if two psychological events (in that case, the prime and target, and in the present case, the prime and the response) are evaluatively congruent with each other (e.g., Wentura, 2000). Indeed, in line with this possibility, participants in both conditions of Experiment 2 were more likely than not to choose the "influenced" rather than the "not influenced" response following congruent rather than incongruent trials (independently of whether the default response was that they were not vs. they were influenced by the prime; 43% vs. 20% in the former and 65% vs. 38% in the latter condition). If influences of response set indeed contributed to responding on the prime influence measure in this and the remaining experiments, this would raise additional serious concerns about its internal validity. We thank an anonymous reviewer for raising this possibility.

its robustness to procedural variations, which is of inherent theoretical interest, and, indirectly, regarding (*b*) the cognitive mechanisms giving rise to self-reports of prime influence. Specifically, on a view under which such self-reports are unadulterated reflections of accurate introspection, response format should have at most a small effect on rates of reported influence awareness; on a view under which such self-reports are the output of post hoc inferential processes, response format could produce sizable effects.

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiments 1–2.

Participants and design

A total of 464 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/implicit/). Participants were excluded from analyses if they did not complete the AMP ($n = 60$) or pressed the same key on all AMP trials ($n = 74$). Given that 18 participants were excluded based on multiple criteria, the final sample size was 348.

Participants represented a total of 28 countries of citizenship, with the United States accounting for most of the sample ($n = 248$, or 71%) and no other country accounting for more than 8%. Eighty-seven percent of participants were from majority-English-speaking countries. In the final sample, 222 participants were women, 117 men, and 7 of other genders. The mean age was 34 years ($SD = 14$).

The design was similar to Experiments 1–2, with the exception that influence awareness was measured differently. Specifically, participants were asked to press one key to indicate that the prime had influenced their response and a different key to indicate that it had not. As such, the asymmetry between go and no-go responses that had characterized Experiments 1–2 was removed. Additionally, as in the previous experiments, prime valence (positive vs. negative) was manipulated within participants, and influence awareness (unaware vs. aware) was measured on each trial.

Procedure and measures

AMP The AMP procedure was identical to the one used in Experiments 1–2, with the exception that participants were asked to press the 1 key on their keyboard if they thought their response to the target had been influenced by the prime and the 0 key if they thought their response to the target had not been influenced by the prime. As such, unlike in previous experiments, no response deadline was imposed; rather, participants were able to advance to the next trial only after responding to the influence awareness item.

Exploratory measures The same exploratory measures were collected as in Experiment 2, with the exception of the items on which participants indicated (*a*) whether their data should or should not be retained for analyses and (*b*) the proportion of trials in which they pressed the space bar to advance more quickly to the next trial. The former measure produced no effect in Experiments 1–2 and, as such, we removed it from all subsequent experiments. The latter measure was irrelevant to the present experiment given that the distinction between go and no-go responses had been eliminated.

Results

The proportion of pleasant responses by prime condition (positive vs. negative) and self-reported influence awareness (unaware vs. aware) is shown in Fig. 3.

In terms of the overall AMP effect, we replicated the results of Experiments 1–2. Specifically, participants were more likely to rate the target as pleasant following positive, $P(\text{pleasant}) = 0.730$, than negative primes, $P(\text{pleasant}) = 0.344$, resulting in a mean AMP effect of $P_{\text{diff}}(\text{pleasant}) = 0.386$ (median = 0.429; $SD = 0.345$). The mean proportion of aware trials at the participant level was $P(\text{aware}) = 0.479$ (median = 0.492; $SD = 0.348$), i.e., halfway between the standard and reversed conditions of Experiment 2. Importantly, as in previous experiments, participants showed a larger AMP effect on aware, $P_{\text{diff}}(\text{pleasant}) = 0.606$, than on unaware trials, $P_{\text{diff}}(\text{pleasant}) = 0.181$. As in Experiments 1–2, an AMP effect was clearly present on unaware trials. Given the roughly equal base rates of aware and unaware trials, the size of the overall AMP effect was halfway between the aware and unaware AMP effects.

Accordingly, the best-fitting model with valence of response (pleasant vs. unpleasant) as the dependent variable contained random intercepts for participants, primes, and targets, as well as a Prime Type (positive vs. negative) × Awareness (unaware vs. aware) interaction as the fixed effect, $\chi^2(1) = 1302.82$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that the strength of the AMP effect differed significantly across the two levels of the awareness variable. Notably, in planned comparisons, we found a significant AMP effect on both unaware, $b = 1.00$, $z = 6.60$, $p < 0.001$, and aware trials, $b = 3.46$, $z = 22.20$, $p < 0.001$.

Discussion

We replicated the awareness effect originally established by Hughes et al. (2023) and also obtained in Experiments 1–2 above. Specifically, the AMP effect was larger on influence-aware than on influence-unaware trials. The base rate of aware trials was about 48%, that is, halfway between the one obtained on the unaware-default AMP (Experiment 1

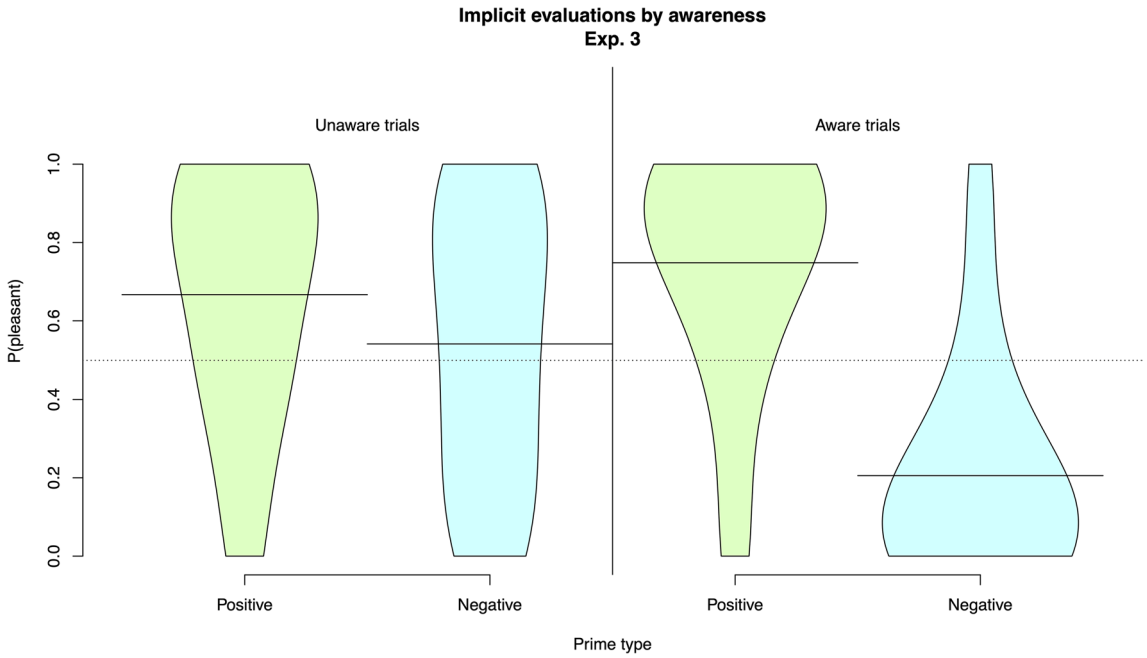

Fig. 3 Implicit (AMP) evaluations by prime type (positive vs. negative) and self-reported influence awareness at the trial level (unaware vs. aware; Experiment 3). The *y*-axis shows the proportion of pleasant responses, with the dashed line at 0.5 marking neutrality. The solid lines represent condition means

and standard condition of Experiment 2) and on the aware-default AMP (reversed condition of Experiment 2). As such, the overall AMP effect was virtually equidistant from the AMP effects that emerged on aware and on unaware trials.

Along with Experiment 2, we take the present data to provide initial evidence for the idea that responses on the prime influence item emerge from informationally promiscuous inferential processes, which strongly reflect (among other factors) the nature of the response options. (The inferential account was tested more directly in Experiments 5–7.) From a practical perspective, when using the asymmetric awareness measure in Experiments 1–2, we observed (*a*) volatility in awareness base rates depending on the default option and (*b*) a correlation between self-reported skipping and the proportion of aware responses. As such, in Experiments 4–6 we rely on the binary measure of influence awareness introduced in this experiment.

Experiment 4

To achieve appropriate experimental control, the modified AMP procedure used by Hughes et al. (2023), and by extension in the present Experiments 1–3, relies exclusively on a set of normatively highly positive and normatively highly negative prime images. However, this aspect of the design represents a substantial departure from the AMP procedure as implemented in most of the relevant literature, where primes tend to be relatively mildly valenced stimuli, such as human faces, members of different racial groups, or words or images related to mental illness. As such, along with Experiments 2–3, the present study provides important evidence on variability in the size of the awareness effect. After all, even assuming that the awareness effect causes the AMP effect (rather than vice versa), much of the purported issue is eliminated if influence awareness and the AMP effect are statistically related to each other only for extreme primes but not (or to a lesser degree) for primes customarily used in AMP research.

Highly valenced stimuli (Theeuwes & Belopolsky, 2012; Wentura et al., 2014), and especially highly negative stimuli (Brosch & Sharma, 2005; Öhman et al., 2001), are known to attract visual attention. Once a valenced stimulus is attended to, it is more likely to enter into explanations of one's behavior, either because the stimulus itself becomes temporarily more salient or because the affective reaction that it generates is stronger. As such, we reasoned that the use of only highly positive and highly negative prime stimuli in Experiments 1–3 may have resulted in inflated estimates of the base rate of aware trials and, crucially, of the relationship between influence awareness and the size of the AMP effect. Initial evidence for this conjecture is provided by Experiment 5 of Hughes et al. (2023), where participants completed both a modified AMP using the normatively positive and negative primes described above and a modified AMP using the faces of Barack Obama and Donald Trump

as prime stimuli. Although the proportion of aware trials and the AMP effect was correlated for both AMPs, the effect was considerably larger for the AMP using evaluatively extreme primes ($r = 0.600$) than for the AMP using non-extreme primes ($r = 0.172$). These two correlations are significantly different from each other, $z = 6.18$, $p < 0.001$.

However, the two AMPs included in Experiment 5 of Hughes et al. (2023) differed from each other in several ways beyond stimulus valence, including the fact that one AMP used two sets of semantically diffuse stimuli and the other AMP used two specific attitude objects as primes. Therefore, making inferences about the source of the discrepancy in the size of the awareness effect in that study is difficult. As such, in the present experiment, we manipulated prime valence continuously rather than relying on only highly valenced primes (as we did in Experiments 1–3) or on two separate sets of primes differing from each other in multiple ways (as Hughes et al. did in their Experiment 5).

Method

Unless otherwise noted, the design, materials, procedure, and analytic strategy were identical to those used in Experiments 1–3.

Participants and design

A total of 545 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/implicit/). Participants were excluded from analyses if they did not complete the AMP ($n = 52$) or pressed the same key on all AMP trials ($n = 93$). Given that 16 participants were excluded based on multiple criteria, the final sample size was 416.

Participants represented a total of 38 countries of citizenship, with the United States accounting for most of the sample ($n = 297$, or 71%) and no other country accounting for more than 5%. Eighty-four percent of participants were from majority-English-speaking countries. In the final sample, 262 participants were women, 145 men, and 8 of other genders. The mean age was 39 years ($SD = 15$).

The design was similar to Experiments 1–3, with the exception that prime valence was manipulated continuously within participants, rather than as a binary variable. Influence awareness was measured at the trial level in the same way as Experiment 3. That is, participants pressed one key to indicate influence and a different key to indicate lack of influence, thus removing the asymmetry between go and no-go responses that had characterized Experiments 1–2.

Materials

We used the same target stimuli as in Experiments 1–3; however, the prime stimuli were different. Specifically, rather than focusing only on extremely positive versus extremely negative images, our goal was to cover as much of the valence space as possible, including mildly valenced and neutral images. To this end, we selected a subset of 200 images from the OASIS stimulus set (Kurdi et al., 2017), which consists of a total of 900 images, using the following procedure.

First, to exclude evaluatively ambiguous stimuli, we restricted the set to images that had a valence standard deviation of 1 or less on a seven-point scale. This restriction resulted in a smaller set of 271 images. Second, we randomly sampled sets of 200 images from this smaller set 10,000 times and chose the subset with the largest amount of variability in valence across images ($SD = 1.48$) to make sure that coverage was as comprehensive as possible. Valence values in this final set of images ranged from 1.11 to 6.49, with a mean of 4.24 and a median of 4.26.

Procedure and measures

The AMP procedure and exploratory measures were identical to the ones administered in Experiment 3, with two exceptions. Given that the number of prime stimuli exceeded the number of AMP trials ($200 > 80$), unlike in previous experiments, each trial featured a unique prime. Moreover, across the 70 experimental trials, primes were sampled under the constraint that half of them were below and half of them above the median valence.

Analytic strategy

As preregistered, we expected the relationship between prime valence and the probability of a pleasant response to be nonlinear. As such, instead of the generalized linear mixed-effects models from Experiments 1–3, here we used a more flexible generalized additive mixed-effects model. These models are implemented in the mcgv package (version 1.8.33; Wood, 2017) in the R statistical computing environment.

Results

The proportion of pleasant responses by prime valence (varying continuously on a 1–7 scale) and self-reported influence awareness (unaware vs. aware), along with predicted values derived from the best-fitting generalized additive mixed-effects model, is shown in Fig. 4.

Before we turn to the interpretation of the model that relies on the continuous valence variable, to ensure comparability with the previous experiments, we first report descriptive statistics using a median split of the valence variable (positive vs. negative). In terms of the overall AMP effect, we replicated the results of Experiments 1–3.

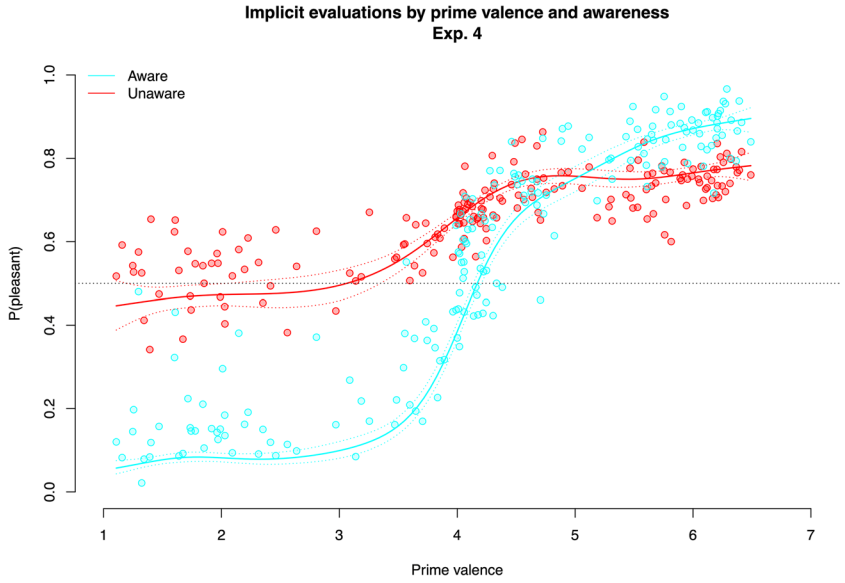

Fig. 4 Implicit (AMP) evaluations by prime valence and self-reported influence awareness at the trial level (unaware vs. aware; Experiment 4). The *y*-axis shows the proportion of pleasant responses, with the dashed line at 0.5 marking neutrality. The dots are observed values, and the lines represent predicted values from the corresponding generalized additive mixed-effects model

Specifically, participants were more likely to rate the target as pleasant following positive, $P(\text{pleasant}) = 0.768$, than negative primes, $P(\text{pleasant}) = 0.475$, resulting in a mean AMP effect of $P_{\text{diff}}(\text{pleasant}) = 0.293$ (median = 0.314; $SD = 0.249$). Unsurprisingly, given the less extreme nature of the primes, the AMP effect was smaller than in previous experiments and closer in size to the AMP effects customarily obtained in the literature.

The mean proportion of aware trials at the participant level was $P(\text{aware}) = 0.404$ (median = 0.371; $SD = 0.329$), i.e., similar but somewhat lower than in Experiment 3. Importantly, as in previous experiments, participants showed a larger AMP effect on aware, $P_{\text{diff}}(\text{pleasant}) = 0.525$, than on unaware trials, $P_{\text{diff}}(\text{pleasant}) = 0.136$. As in previous experiments, an AMP effect was present on unaware trials. Given that the base rate of aware trials decreased from Experiment 3, the size of the overall AMP effect was again more similar to the unaware than to the aware AMP effect.

The best-fitting model with valence of response (pleasant vs. unpleasant) as the dependent variable contained a random intercept for participants and an interaction between a spline for prime type (varied continuously) and the binary awareness variable (unaware vs. aware), $\chi^2(6.18) = 104.17$, $p < 0.001$, for unaware responses, and $\chi^2(7.44) = 486.10$, $p < 0.001$, for aware responses. In the preregistration, we had also specified random intercepts for primes and targets; however, this model was too complex to fit. As such, we chose to include only a random intercept for participants given that the participant level consistently accounted for the largest amount of random-effects variance in Experiments 1–3. Because coefficients obtained from generalized additive mixed-effects models lack intuitive interpretations, below we rely on a visual inspection of the line of best fit.

The inspection of fitted values shown in Fig. 4 suggests four main conclusions. The first two conclusions represent replications of previous findings, whereas the latter two concern more fine-grained inferences about the size of the awareness effect across the entirety of the valence spectrum.

First, as in previous experiments, an AMP effect was apparent on both aware and unaware trials, as indicated by the difference in the probability of a pleasant response at the minimum and maximum values of the valence variable. Second, the awareness effect obtained in previous experiments also emerged in these data, with a noticeably larger AMP effect on aware than on unaware trials.

Third, we newly obtained an extremity effect such that the discrepancy in AMP effects between aware and unaware trials was most pronounced for highly negative, and to some degree highly positive, primes and less pronounced for more moderately valenced primes. Fourth, the data were characterized by an even stronger valence asymmetry effect such that the awareness effect was particularly large for highly negative stimuli (below a valence of 3), moderate for moderately negative stimuli (between 3 and 4), and small or nonexistent for positive stimuli (above 4). Although the granularity of this pattern is unique to Experiment 4, an inspection of Figs. 1, 2, 3 shows that the awareness effect was also consistently stronger for negative than positive stimuli in the data obtained from Experiments 1–3.

Discussion

In line with our expectations, we found an overall decrease in the proportion of aware trials relative to Experiment 3, which had relied exclusively on highly valenced primes. Therefore, in this experiment, the total AMP effect was more similar to the AMP effect emerging on unaware trials than to the AMP effect emerging on aware trials. Crucially, we also obtained a strong valence extremity effect such that the difference between aware and unaware AMP effects was large on trials involving extreme (and especially extremely negative) primes and considerably smaller or nonexistent on trials involving moderately valenced and even extremely positive primes.

As such, the results of this experiment suggest that Hughes et al. (2023) and, by extension, Experiments 1–3 above may well have overestimated the extent of the awareness effect given their exclusive reliance on highly valenced primes. Notably, it seems reasonable to assume that the size of the association between AMP effects and self-reports of influence awareness tends to be modest in standard AMP studies given that such studies generally feature moderately valenced, rather than extremely valenced, primes.

Experiment 5

In Experiments 5–7, we turned to more directly investigating the cognitive mechanisms undergirding self-reports of prime influence. Specifically, Hughes et al.'s (2023) conclusion about the awareness effect being causally responsible for the AMP effect would seem more defensible if evidence could be marshaled against the reverse causal direction. As such, in Experiments 5–7, we attempted to vindicate Hughes et al. (2023) by ruling out the inferential account of their findings. To preview the present results: the inferential account could not be eliminated. On the contrary, the data suggest that the inferential account of the correlation between AMP effects and awareness reports is eminently plausible. As such, although the present data cannot be used to conclusively arbitrate between the privileged introspective access view and the inferential view, the claim that AMP effects depend on awareness has no more support now than it did prior to the publication of Hughes et al. (2023).

In Experiment 5—the initial experiment in the set of studies investigating how self-reports of prime influence emerge—we asked participants to provide judgments of influence awareness not only with respect to their own AMP performance (as in Experiments 1–4) but also, additionally, with respect to a past participant's AMP performance from Experiment 3. We reasoned that participants may rely on a lay theory about evaluative congruency between the prime and the response to make judgments of prime influence, specifically that participants may be more likely to report influence awareness for positive prime–pleasant response and negative prime–unpleasant response combinations than for positive prime–unpleasant response and negative prime–pleasant response combinations (the "inferential account"). If participants possess a lay theory of this kind, then (*a*) the patterns of first-person and third-party influence awareness judgments should be similar to each other and (*b*) first-person and third-party influence awareness rates should be correlated with each other at the participant level.

Crucially, a relationship between third-party influence awareness judgments and AMP performance, by definition, cannot emerge from the former causing the latter. Moreover, a correlation between first-person and third-party influence awareness rates may be indicative of the two types of judgment relying on shared processes. Of course, the present data cannot be used to eliminate the possibility of the causal direction suggested by Hughes et al. (2023). Nonetheless, unlike the purely correlational results presented by those authors, the present findings have the ability to provide direct evidence for a particular causal direction.

Method

Unless otherwise noted, the design, materials, and procedure were identical to those used in Experiment 3.

Participants and design

A total of 367 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/implicit/). Participants were excluded from analyses if they did not complete the first-person AMP ($n = 22$) or the third-party AMP ($n = 9$), or pressed the same key on all first-person AMP trials ($n = 54$). Given that eight participants were excluded based on multiple criteria, the final sample size was 292.

Participants represented a total of 29 countries of citizenship, with the United States accounting for most of the sample ($n = 194$, or 67%) and no other country accounting for more than 8%. Seventy-eight percent of participants were from majority-English-speaking countries. In the final sample, 160 participants were women, 92 men, and 12 of other genders. The mean age was 32 years ($SD = 15$).

The design was similar to the design of Experiment 3, with the crucial deviation that each participant completed both a first-person AMP (identical to the AMP in Experiment 3) and a third-party AMP in which they made influence

judgments with respect to the AMP responses of a past participant rather than with respect to themselves. As such, in addition to the manipulated within-participant variable of prime valence (positive vs. negative) and measured within-participant variable of awareness (unaware vs. aware), we also manipulated AMP type (first-person vs. third-party AMP) within participants.

Procedure and measures

Familiarization with primes and targets Given that some participants in Experiments 1–4 had expressed confusion about the difference between the target images (abstract paintings) and the noise mask, before the AMP, participants were exposed to a table featuring a randomly selected subset of six (three positive and three negative) primes and six targets.

AMP The AMP procedure was similar to the AMP procedure of Experiment 3, with the crucial deviation that participants completed both a first-person AMP and a third-party AMP. The first-person AMP was procedurally identical to the AMP implemented in Experiment 3. That is, participants completed AMP trials consisting of highly positive or highly negative primes and evaluatively ambiguous abstract painting targets and then were asked to judge whether their response to the target was influenced by the prime. On the newly added third-party AMP, participants observed the same prime–target combinations as a past participant. However, instead of responding to the target themselves, they were exposed to the past participant's response and then were asked to judge whether the past participant (rather than they themselves) had been influenced by the prime.

For the purposes of the experiment, each participant was yoked to one of 296 participants from Experiment 3. These participants were selected under the constraint that they had to exhibit nonzero variance both on their AMP responses (pleasant vs. unpleasant) and on their awareness responses (unaware vs. aware). For each new participant, a specific past participant was randomly selected at the outset of the experiment.

The randomly selected past participant's experience and responses in the experiment determined the present participant's experience in the following way. The past participant's 10 practice trials (i.e., prime–target combinations) served as the new participant's practice trials. The order of the practice trials was randomized anew. Crucially, of the past participant's remaining 70 trials, 34 randomly selected ones (17 with positive primes and 17 with negative primes) served as trials on the first-person AMP and the remaining 36 trials (18 with positive primes and 18 with negative primes) as trials on the third-party AMP. All participants completed the first-person AMP first and the third-party AMP second. The instructions for influence awareness trials on the third-party AMP were modeled after the instructions for the first-person AMP, with the exception that they referred to the past participant's responses rather than the participant's own responses.

As mentioned above, the practice trials and first-person AMP trials were procedurally identical to the AMP trials in Experiment 3. On third-party AMP trials, the participant was exposed to the prime–target combination from the past participant's trial (under identical timing conditions). To proceed from the noise mask, instead of entering their own response, the participant was asked to press the space bar. Once the space bar had been pressed, the participant was exposed to a screen that read, "The [past] participant's response on this trial was: Pleasant/Unpleasant." Similar to the first-person AMP, the participant was asked to press the 1 key if they thought that the past participant's response to the target was influenced by the prime and to press 0 if they thought that it was not. Once this judgment was entered, the program moved on to the next trial. No response deadline was imposed on awareness judgments on either AMP.

Exploratory measures The exploratory measures were similar to the ones collected in Experiments 1–4, but they took into account the fact that participants completed both a first-person and a third-party AMP. Specifically, participants were asked to estimate the proportion of trials in which (*a*) their own responses were influenced by the prime and (*b*) the past participant's responses were influenced by the prime; to describe, in a free response format, the strategy that they used to decide whether (*c*) their own responses and (*d*) the past participant's responses were influenced by the prime and to (*e*) reflect on whether the two strategies were similar or different; and, finally, (*f*) to share any other comments that they might have on the experiment.

Results

The proportion of aware responses by AMP condition (first-person vs. third-party), prime condition (positive vs. negative), and response (pleasant vs. unpleasant) is shown in Fig. 5. In line with our shift in emphasis from Experiments 1–4, in this and the two remaining experiments we focused on the influence awareness variable, rather than the AMP response, as the dependent measure.

Both the first-person AMP, $P_{\text{diff}}(\text{pleasant}) = 0.350$ (median = 0.353; $SD = 0.358$), and the third-party AMP, $P_{\text{diff}}(\text{pleasant}) = 0.348$ (median = 0.389; $SD = 0.380$), were characterized by an overall AMP effect. This finding is theoretically inconsequential but is a precondition for meaningful analyses of the awareness items. More importantly, the AMP effect was indistinguishable across the two AMP versions,

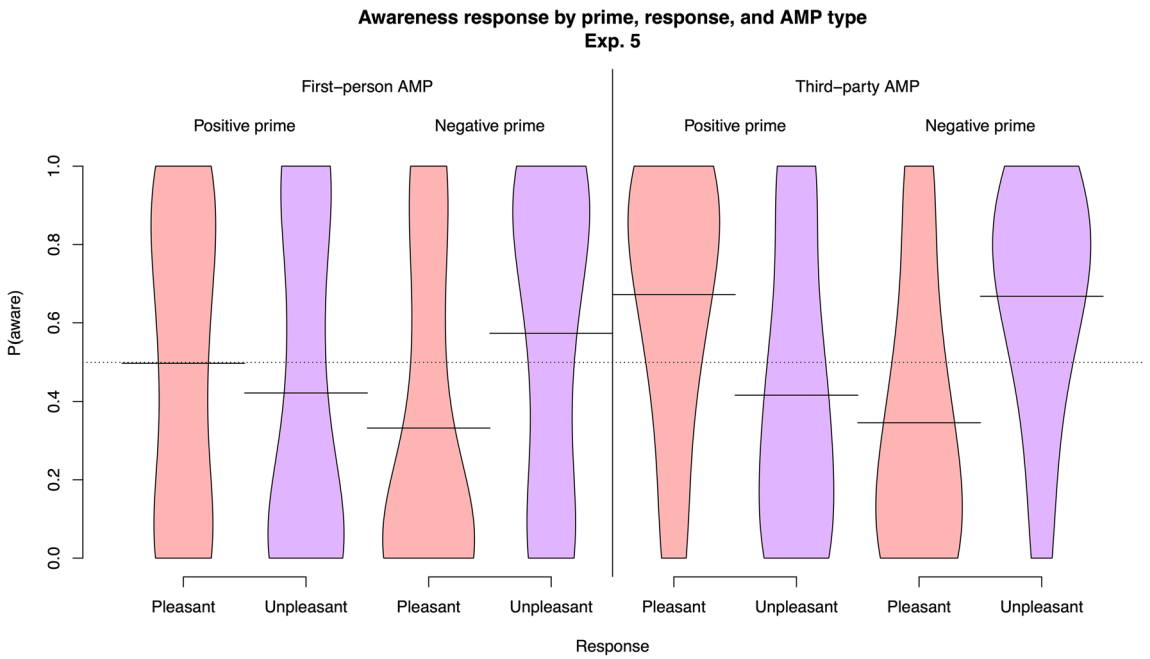

Fig. 5 Influence awareness responses by AMP type (first-person vs. third-party), prime type (positive vs. negative), and trial response (pleasant vs. unpleasant; Experiment 5). The *y*-axis shows the proportion of aware responses, with the dashed line at 0.5 marking equiprobability. The solid lines represent condition means

suggesting that any differences in awareness effects cannot be accounted for by differences in AMP performance alone.

The first-person AMP was characterized by somewhat lower overall influence awareness rates, $P(\text{aware}) = 0.477$ (median = 0.471; $SD = 0.342$) than the third-party AMP, $P(\text{aware}) = 0.566$ (median = 0.583; $SD = 0.250$). On the first-person AMP we replicated the congruency effect already observed in Experiments 1–4. Specifically, participants were more likely to report influence awareness following pleasant, $P(\text{aware}) = 0.512$, than unpleasant, $P(\text{aware}) = 0.395$, responses when primes were positive. Conversely, they were more likely to report influence awareness following unpleasant, $P(\text{aware}) = 0.594$, than pleasant, $P(\text{aware}) = 0.272$, responses when primes were negative.

Crucially from a theoretical perspective, a similar congruency effect emerged on the third-party AMP. Specifically, just as on the first-person AMP, participants were more likely to report influence awareness following pleasant, $P(\text{aware}) = 0.679$, than unpleasant, $P(\text{aware}) = 0.345$, responses when primes were positive. Conversely, they were more likely to report influence awareness following unpleasant, $P(\text{aware}) = 0.676$, than pleasant, $P(\text{aware}) = 0.325$, responses when primes were negative. Notably, the awareness effect was larger and more symmetric across prime valences on the third-party AMP than on the first-person AMP.

Accordingly, the best-fitting model with influence awareness (aware vs. unaware) as the dependent variable contained random intercepts for participants and primes, as well as a Prime Type (positive vs. negative) × Response (pleasant vs. unpleasant) × AMP Type (first-person vs. third-party) interaction as the fixed effect, $\chi^2(3) = 139.64$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that the congruency effect was present on both the first-person and the third-party AMP, but stronger on the latter than on the former. In planned comparisons, we found a significant congruency effect following both positive primes, $b = 0.41$, $z = 5.05$, $p < 0.001$, and negative primes, $b = 1.33$, $z = 16.62$, $p < 0.001$, on the first-person AMP. Crucially, similar but larger congruency effects were also obtained following both positive primes, $b = 1.87$, $z = 23.76$, $p < 0.001$, and negative primes, $b = 1.63$, $z = 19.93$, $p < 0.001$, on the third-party AMP.

In a final preregistered analysis, we probed whether reported rates of influence awareness on the first-person and third-party AMPs were associated with each other. Such a relationship would provide initial evidence for the idea that influence awareness judgments for the self and third parties operate based on a shared mechanism not merely at the group level but even at the level of individual raters. In line with this idea, we found a robust correlation between the proportion of aware judgments for the self and the past participant at the individual level, $r = 0.487$, $t(263) = 9.04$, $p < 0.001$. In other words, participants who believed that they themselves were often influenced by the primes tended to believe that the past participant whose performance they were asked to judge had also often been influenced by the primes and vice versa. This association cannot be explained

by any similarities in AMP effects across present and past participants given successful random assignment, $r = 0.007$, $t(263) = 0.11$, $p = 0.912$, $BF_{01} = 6.94$.

Discussion

In this experiment, we replicated the congruency effect obtained in Experiments 1–4 such that participants were more likely to report being influence-aware when the prime and the response were evaluatively congruent with each other than when they were not. Crucially, however, the congruency effect emerged not only when participants made influence awareness judgments with respect to their own performance but even with respect to a past participant's performance. This pattern of results is obviously incompatible with the possibility that third-party awareness judgments could have influenced a past participant's AMP performance and therefore provides compelling evidence for the alternative causal direction.

Of course, the present results do not conclusively rule out the introspective view favored by Hughes et al. (2023); however, at a minimum, they suggest that an alternative causal account is plausible. Nonetheless, proponents of the introspective access account may argue that first-person and third-party influence awareness judgments rely on different mechanisms. Indeed, the correlational evidence is not conclusive regarding the underlying cognitive mechanism. However, given the robust correlation that we obtained between first-person and third-party judgments of awareness, an interpretation assuming a common cause provides a parsimonious and complete account of these data. As such, if Hughes et al. (2023) wish to continue to advance an account about awareness of prime influence causing AMP effects, it is up to them to provide some explanation consistent with that account of why first-person and third-party judgments of influence are robustly associated with each other at the participant level.

Notably, Hughes et al. (2023) found that participants' influence awareness judgments from one AMP were highly correlated with their performance on a different AMP. According to these authors, this result raises the possibility that participants are characterized by some unknown individual difference that makes them relatively more or less influence-aware, and this individual difference (rather than construct-relevant variance in implicit evaluations), in turn, drives responding across all AMPs. The present results offer a different explanation for these findings: Hughes et al. (2023) may have obtained the cross-AMP correlations because participants differ in terms of the extent to which they subscribe to the lay theory about prime–response congruency; therefore, judgments of influence awareness will be correlated across AMPs without necessarily contributing to AMP performance. Notably, in the present experiment, first-person and third-party awareness were correlated with each other although, by design, first-person and third-party AMP scores were unrelated.[12]

Experiment 6

In Experiment 5, we found similar patterns of awareness judgments irrespective of whether those judgments were made with respect to the self or a past participant. This finding suggests that both types of judgment may rely on a shared mechanism, presumably a lay theory about evaluative congruency between primes and responses. This possibility is bolstered by the result that first-person and third-party

[12] Another piece of evidence against Hughes et al.'s (2023) account emerges from a recent study by Morris and Kurdi (2023). In this study, each participant completed five AMPs randomly selected from a larger set of 16 adapted from Nosek (2005). Crucially, the attitude objects were diverse and included comparisons such as American vs. Canadian, cats vs. dogs, Coke vs. Pepsi, thin people vs. fat people, and Yankees vs. Diamondbacks. As such, there is no theoretical reason to expect these attitudes, as a set, to be correlated with each other; rather, any high intercorrelation may be seen as evidence for method-specific variance, perhaps of the kind suggested by Hughes et al. (2023), inflating the statistical relationship. However, in fact, the average correlation across different AMPs was $r = -.001$ and did not differ significantly from zero, $t(119) = -0.30$, $p = .765$, $BF_{01} = 9.44$, Cohen's $d = -0.03$, or from the correlation among explicit evaluations toward the same comparisons, $t(119) = -0.86$, $p = .394$, $BF_{01} = 6.21$, Cohen's $d = -0.08$.

Hussey and Cummins (2022) disputed the validity of this interpretation, pointing out that if the absolute deviation from neutrality in AMP performance is used as the dependent variable in this analysis, an ICC of .26 is observed, suggesting shared variance across different AMPs. Even putting theoretical considerations about the accuracy of relying on absolute deviations aside, it is unclear how this intercorrelation provides evidence for Hughes et al.'s (2023) account given that scores on different AMPs may be correlated with each other for a host of different reasons having nothing to do with awareness of prime influence.

In fact, the correlation between the participant-level mean absolute deviation from neutrality on the set of five AMPs and mean absolute deviation from neutrality on the parallel set of five explicit measures in this sample was $r = .464$, $t(566) = 12.47$, $p < .001$. Importantly, a mean correlation of $r = .170$ was also observed when the participant-level correlation between extremity in implicit evaluations and extremity in explicit evaluations was calculated using nonoverlapping subsets of attitude objects, thus suggesting that the relationship is not entirely due to variance shared between implicit and explicit evaluations of the same targets. This result, although not conclusive, raises the possibility that the correlation observed by Hussey and Cummins (2022) in absolute deviation across different AMPs is a result of an individual difference having to do with broader tendencies in evaluative behavior, such as the need to evaluate (Jarvis & Petty, 1996), rather than any specific aspect(s) of AMP performance. And, of course, this is only one of many potential hypotheses about why such intercorrelations might emerge.

awareness judgments were robustly correlated with each other at the participant level.

The goal of the present experiment was to provide additional, and more direct, evidence for the account that first-person judgments of influence awareness can rely on an inferential mechanism rather than privileged introspective access to the source of one's evaluations. (Similar to Experiment 5, this experiment was not designed to rule out the introspective view favored by Hughes et al., 2023; rather, the goal was to provide evidence for the plausibility of an alternative causal mechanism.) To this end, we implemented a modified AMP procedure in which the main trials did not involve any presentation of primes. Rather, participants were misled to believe that primes had been presented subliminally and made judgments of influence awareness based on an ostensible post hoc prime that was actually presented for the first time as part of the awareness judgment (for a similar design involving no actual presentation of prime stimuli, see Ruys et al., 2012). As such, as in Experiment 5, influence awareness could not have contributed to the AMP effect given that, in the absence of primes, no AMP effect could have emerged in the first place.

Method

Unless otherwise noted, the design, materials, and procedure were identical to those used in Experiment 3.

Participants and design

A total of 206 participants were recruited from the Prolific Academic crowdsourcing website (https://www.prolific.co/) and paid $1.50 in exchange for their participation. We were unable to recruit participants from Project Implicit for this experiment given that Project Implicit does not allow for studies involving deception. Participants were excluded from analyses if they did not complete the AMP ($n = 2$) or pressed the same key on all AMP trials ($n = 19$), resulting in a final sample size of 185. All participants were recruited from the United States. In the final sample, 89 participants were men, 88 women, and one non-binary. The mean age was 38 years ($SD = 13$).

The design was identical to Experiment 3. That is, participants completed the AMP and, following each trial, provided a judgment about whether their response was influenced by the prime or not. Prime valence (positive vs. negative) was manipulated within participants and awareness (unaware vs. aware) measured on each trial. The sole major deviation from Experiment 3, described in more detail below, was that the AMP trials did not include any actual prime stimuli. Instead, ostensible primes were presented only post hoc, after the participant had already entered their AMP response.

Procedure and measures

Prior to the AMP, participants were informed that the primes would be presented so quickly that they might not be visible to the naked eye but might still bias responses to the targets (see Ruys et al., 2012). Unlike in Experiment 3, and similar to Experiment 5, participants were exposed to a randomly selected subset of six primes and six targets before completing the AMP. Notably, both in this phase and during the AMP, primes were displayed in grayscale to make the cover story more believable.

The AMP procedure was identical to the one used in Experiment 3, with one crucial exception. Specifically, AMP trials themselves did not feature a prime stimulus; instead, a light gray square was programmed to be displayed before the target for a duration of 16 ms.[13] Based on the open-ended responses to the relevant exploratory items, some participants perceived this stimulus as a brief flicker, whereas other did not consciously perceive it at all. After the participant had responded to the target, they were asked to indicate whether their response was influenced by the prime. The ostensible prime was displayed in grayscale below the sentence, "[t]his is the real-life image that was displayed before the abstract painting." As in Experiment 3, participants were asked to use the 1 key to indicate influence awareness and the 0 key to indicate lack of awareness.

At the end of the experiment, following the same set of exploratory items used in Experiment 3, participants completed standard demographic variables, including age, gender, sexual orientation, race, education, income, and political ideology.

Results

The proportion of aware responses by prime condition (positive vs. negative) and response (pleasant vs. unpleasant) is shown in Fig. 6.

Unlike in previous experiments, no overall AMP effect was observed given that primes were presented post hoc and therefore could not have influenced responding to the targets, $P_{diff}(\text{pleasant}) = -0.005$ (median $= -0.003$; $SD = 0.116$). Reported awareness rates were lower than in any previous experiment but were nonzero, $P(\text{aware}) = 0.093$ (median $= 0.000$; $SD = 0.186$), $t(184) = 6.80$, $p < 0.001$, Cohen's $d = 0.50$. The major decrease in awareness rates is unsurprising given that no actual primes were presented on

[13] Depending on participants' monitor settings (specifically, the refresh rates used), the actual duration of stimulus presentation may have differed from 16 ms. However, none of the conclusions of the present experiment depend on the specific length of exposure.

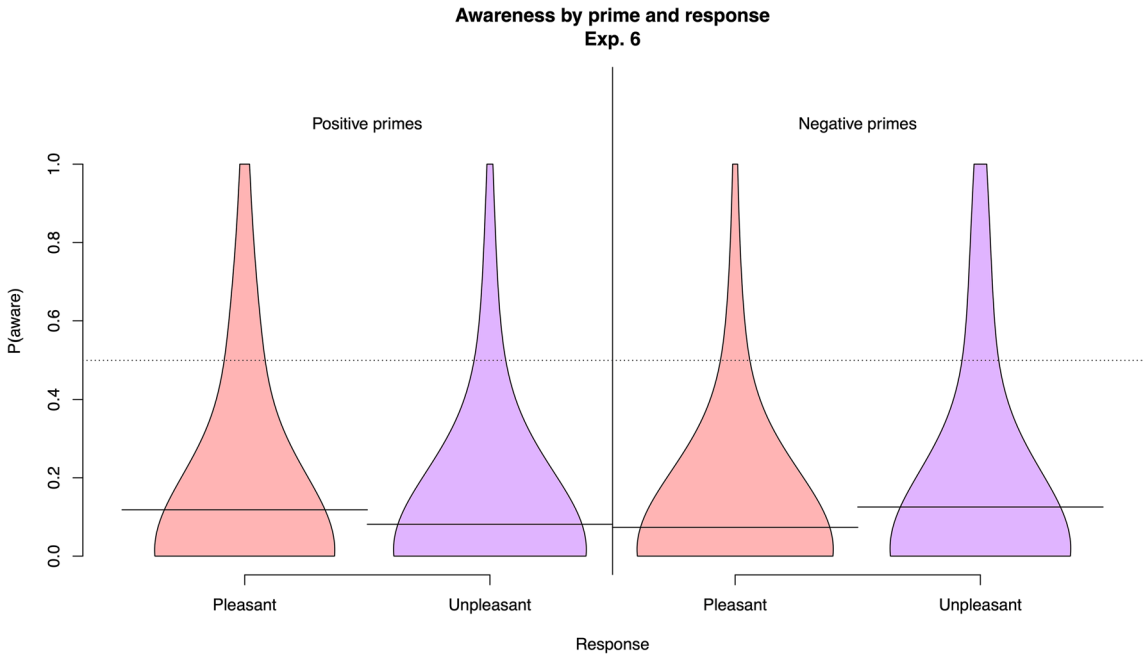

Fig. 6 Influence awareness responses by post hoc prime type (positive vs. negative) and trial response (pleasant vs. unpleasant; Experiment 6). The *y*-axis shows the proportion of aware responses, with the dashed line at 0.5 marking equiprobability. The solid lines represent condition means

the AMP. Critically, nonzero reported awareness rates make meaningful analyses of the awareness measure possible.

Crucially from a theoretical perspective, the congruency effect obtained in all previous experiments was also observed here. Specifically, participants were more likely to report influence awareness following pleasant, $P(\text{aware}) = 0.107$, than unpleasant, $P(\text{aware}) = 0.086$, responses when post hoc primes were positive. Conversely, they were more likely to report influence awareness following unpleasant, $P(\text{aware}) = 0.118$, than pleasant, $P(\text{aware}) = 0.065$, responses when post hoc primes were negative. In other words, evaluative congruency between the post hoc prime and the response was associated with a 25% increase in the probability of an aware response following positive post hoc primes and an 83% increase following negative post hoc primes. This finding is remarkable given that no actual primes were presented on any of the AMP trials.

Accordingly, the best-fitting model with influence awareness (aware vs. unaware) as the dependent variable contained random intercepts for participants and a Prime Type (positive vs. negative) × Response (pleasant vs. unpleasant) interaction as the fixed effect, $\chi^2(3) = 97.05$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that the congruency effect was present following both positive and negative post hoc primes but was more pronounced in the latter than in the former case. Accordingly, in planned comparisons, we found a significant congruency effect following both positive primes, $b = 0.81$, $z = 6.46$, $p < 0.001$, and negative primes, $b = 0.76$, $z = 6.78$, $p < 0.001$.

Discussion

In this experiment, participants were again more likely to report being influence-aware when primes and responses were evaluatively congruent with each other. However, this time, AMP trials did not include any actual prime stimuli and, as such, the congruency effect between the ostensible primes and responses emerged in the absence of a genuine AMP effect. Accordingly, the present results offer particularly strong support for the idea that, on a standard AMP, the AMP effect may well cause influence awareness.

Similar to Experiment 5, the present results do not conclusively eliminate the possibility of the reverse causal direction. However, unlike the studies conducted by Hughes et al. (2023), they do provide causal (rather than merely correlational) evidence for the possibility of AMP effects giving rise to awareness effects rather than vice versa. In addition, the present data further bolster the claim that participants may infer influence awareness from externally observable cues rather than from internally attending to their own mental processes.

Experiment 7

Hughes et al. (2023) have argued that AMP effects are "heavily dependent on influence awareness" (p. 1573). In Experiments 5–6, we have demonstrated that the pattern of data obtained by Hughes et al. (2023) and in the present Experiments 1–4 is compatible with a different possibility also briefly mentioned but not tested by Hughes and colleagues, namely that AMP effects cause participants to self-report influence awareness rather than vice versa. Specifically, we have found that (*a*) an awareness effect was present when participants judged influence awareness with respect to a past participant rather than themselves; (*b*) first-person and third-party judgments of awareness were robustly correlated with each other; and (*c*) the awareness effect does not require an AMP effect to emerge.

In the final experiment, we address a particular claim by Hughes et al. (2023) about the validity of the AMP as a measure of implicit evaluations. Namely, these authors suggest that the awareness effect is incompatible with the AMP effect emerging from an unintentional misattribution mechanism, which was originally proposed by Payne et al. (2005) and for which further evidence was provided across several follow-up experiments (Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2012). But, as we argued in the introduction, even accurate introspective access to (the source of) one's evaluative knowledge is not incompatible with the possibility of evaluative knowledge being retrieved unintentionally. To return to the example from the introduction: One may exhibit an unintentional positive evaluation of an éclair spotted in a patisserie window even if one is or becomes aware of the underlying evaluative response and its source.[14]

However, as mentioned earlier, AMP trials are more ambiguous than the toy situation described above given that the evaluative response may be attributed to one of two stimuli. As such, even if one becomes aware of one source of influence, such awareness does not eliminate the possibility that the other stimulus may also have contributed. As explained above, these inferences are necessarily probabilistic given that the person is not privy to how they would have responded in the absence of the prime. In line with these ideas, Oikawa and colleagues demonstrated that becoming aware of the influence of the prime did not eradicate AMP effects as long as simultaneous influence of the target remained a reasonable possibility and, therefore, the ambiguity about the source of the evaluative response was not fully removed (Oikawa et al., 2011; Ruys et al., 2012). In fact, Hughes et al. (2023) themselves recognize this possibility when stating that, even if they believe that they have been influenced by the prime, participants may "[…] also hold the belief that the target really does have a particular valence" (p. 1580), but then proceed to dismiss misattribution as a likely mechanism accounting for AMP effects.

Driven by these considerations, in the final experiment we sought to probe whether the congruency effect documented by Hughes et al. (2023) and in the present Experiments 1–6 is compatible with an unintentional misattribution mechanism. To this end, we manipulated the evaluative strength of primes within participants. In one condition, we asked participants to report to what extent aspects of the prime contributed to their response, as we had in all previous experiments. In this condition, we expected that the self-reports of influence awareness would track the strength of the primes.

In a newly added, and theoretically crucial, condition, participants were asked to provide judgments of influence with respect to aspects of the target rather than the prime. If, as proposed by Payne et al. (2005), the AMP effect emerges from unintentional misattribution, prime strength should systematically influence judgments of both prime and target influence. By contrast, if, as proposed by Hughes et al. (2023), "[…] misattribution is not the best mechanism to explain [their] AMP effects" (p. 1580), then prime strength should cause increases in judgments of prime influence but not judgments of target influence.

Method

Unless otherwise noted, the design, materials, and procedure were identical to those used in Experiment 3.

Participants and design

A total of 349 participants were recruited from the Amazon Mechanical Turk crowdsourcing website (https://www.mturk.com/) and paid $1.00 in exchange for their participation. Participants were excluded from analyses if they (*a*) pressed the same key on all AMP trials ($n = 28$), (*b*) reported being able to read Chinese and were therefore able to understand the target stimuli (see below; $n = 4$), or (*c*) failed an attention check item at the end of the experiment ($n = 11$). Given that four participants were excluded based on multiple criteria, the final sample size was 310. All participants were recruited from the United States. In the final sample, 158 participants were women, 140 men, five non-binary, and one a trans woman. The mean age was 43 years ($SD = 13$).

[14] What is more, based on Experiments 5–6 it seems that, on the AMP, the source of such awareness is more likely to be a configuration of externally observable events (e.g., the combination of seeing the éclair and an involuntary approach movement toward the bakery) rather than privileged introspective awareness only available to the self. As such, even if intentionality were incompatible with this type of genuine introspective awareness, as yet no compelling evidence has been provided that AMP effects are subject to this type of awareness.

The design was similar to the design of Experiment 3. That is, participants completed an AMP with a within-participant manipulation of valence and awareness measured on each trial. However, unlike in Experiment 3, the valence variable had three levels (neutral, mildly negative, and extremely negative). We focused on negative primes because the awareness effect was more pronounced in the negative domain in Experiments 1–4. In addition, we manipulated between participants whether participants rated the influence of the prime (as they had in all previous experiments) or, newly, the influence of the pleasant and unpleasant aspects of the target on their response. A total of 165 participants were assigned to the prime influence condition and 145 participants to the target influence condition.

Materials

Valenced images A total of 30 valenced images each were obtained from the International Affective Picture System (IAPS; Lang et al., 2008) and used as primes on the AMP: (*a*) 10 neutral images (valence mean = 5.28, *SD* = 0.81), (*b*) 10 mildly negative images (valence mean = 3.27, *SD* = 0.59), and (*c*) 10 extremely negative images (valence mean = 2.78, *SD* = 0.62).

Chinese ideographs Similar to Hughes et al. (2023), we used Chinese ideographs as target stimuli on the AMP. Specifically, the set of 72 Chinese ideographs used by Hughes et al. (2023) was complemented with eight additional images to create a final set of 80.

Procedure and measures

The AMP procedure was similar to the AMP procedure implemented in Experiment 3, with the following exceptions: (*a*) instead of 80 trials, the AMP consisted of 33 trials, of which three were practice trials and 30 were experimental trials; (*b*) instead of 100 ms, primes were presented for a duration of 175 ms; (*c*) instead of 100 ms, the blank screen between the prime and the target was presented for 125 ms; and, finally, (*d*) instead of the "E" and "I" keys, participants used the "Q" and "R" keys to enter their responses. Both primes and targets were sampled randomly, without replacement, resulting in combinations of a unique prime and a unique target on each trial.

Following each trial, participants were asked to judge how much certain aspects of the trial influenced their responding. Participants assigned to the prime influence condition rated the influence of the prime on a 1–4 scale labeled "not at all," "somewhat," "a lot," and "entirely" using their mouse. Participants assigned to the target influence condition, which was newly included in this experiment, rated how much unpleasant and pleasant features of the target influenced their response (in individually randomized order) using the same rating scale. Participants were asked to rate the influence of unpleasant and pleasant features separately to avoid ceiling effects and to provide evidence for discriminant validity of the awareness measure. Similar to Experiment 3, no response deadline was imposed on the influence ratings.

Unlike in previous experiments, no exploratory measures were collected. Following the AMP, participants (*a*) indicated whether they were able to read Chinese; reported their (*b*) age, (*c*) race, and (*d*) gender; and (*e*) completed an attention check.

Analytic strategy

In the preregistration, we specified an ANOVA approach using participant-level means as the main analytic strategy. However, to maintain consistency with the remaining experiments, in the main text we report linear mixed-effects models with awareness as the dependent variable and prime type (neutral, mildly negative, and extremely negative) and rating condition (prime, unpleasant target, and pleasant target) as the predictors. Analyses using the ANOVA approach are available in the open data and yielded the same conclusions as the ones presented here.

Results

The distribution of awareness responses by rating condition (prime, unpleasant target, and pleasant target) and prime condition (neutral, mildly negative, and extremely negative) is shown in Fig. 7.

Overall, participants exhibited an AMP effect such that they were most likely to rate the targets negatively after extremely negative primes, $P(\text{negative}) = 0.596$ ($SD = 0.491$), followed by mildly negative primes, $P(\text{negative}) = 0.460$ ($SD = 0.498$), and finally by neutral primes, $P(\text{negative}) = 0.275$ ($SD = 0.447$). Moreover, participants in the prime influence condition reported the lowest rates of influence awareness following neutral primes (mean = 1.55, *SD* = 0.90), higher rates of influence awareness following mildly negative primes (mean = 1.66, *SD* = 0.97), and the highest rates of influence awareness following extremely negative primes (mean = 1.78, *SD* = 1.04). These results were expected and serve as evidence for the construct validity of the valence manipulation.

We obtained the same pattern of results in the theoretically crucial unpleasant target influence condition and the converse pattern of results in the theoretically crucial pleasant target influence condition. Notably, this pattern emerged although, given random assignment, the targets presented across the three conditions did not systematically differ from each other. Specifically, participants reported the lowest levels of unpleasant target influence following neutral primes (mean = 1.56, *SD* = 0.84), higher rates of unpleasant target influence following mildly negative

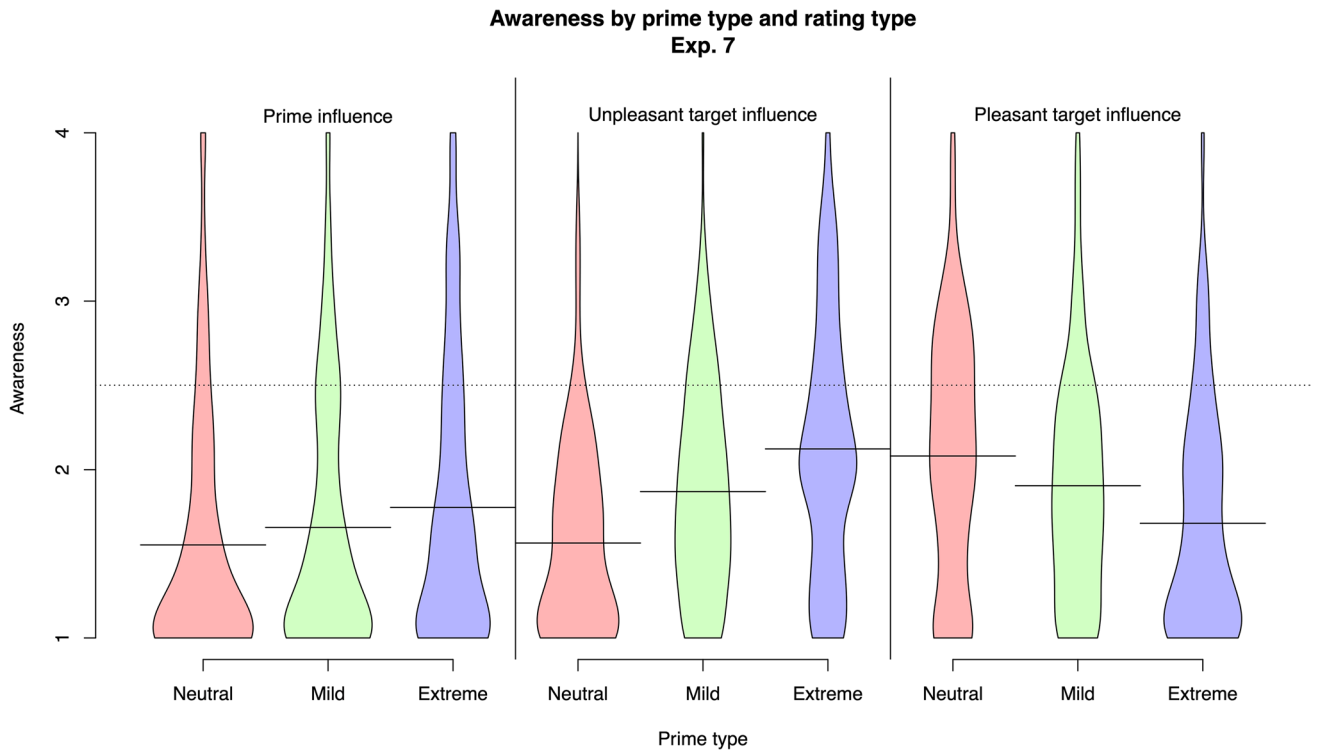

Fig. 7 Ratings of prime influence, unpleasant target influence, and pleasant target influence by prime type (neutral, mildly negative, and extremely negative; Experiment 7). The *y*-axis shows aware responses on a 1–4 scale, with the dashed line marking the scale midpoint. The solid lines represent condition means

primes (mean = 1.87, $SD = 1.01$), and the highest rates of unpleasant target influence following extremely negative primes (mean = 2.12, $SD = 1.07$). Conversely, the highest levels of pleasant target influence were reported after neutral primes (mean = 2.08, $SD = 1.04$), followed by mildly negative primes (mean = 1.90, $SD = 1.02$), and, finally, extremely negative primes (mean = 1.68, $SD = 0.92$).

Accordingly, the best-fitting model with influence awareness as the continuous dependent variable contained random intercepts for participants and primes, as well as a Prime Type (neutral, mildly negative, and extremely negative) × Rating Condition (prime, unpleasant target, and pleasant target) interaction as the fixed effect, $\chi^2(4) = 548.11$, $p < 0.001$. Overall, in line with the descriptive statistics reported above, the best-fitting model suggests that reports of influence increased with increasing prime strength in the prime and unpleasant target rating conditions and decreased with increasing prime strength in the pleasant target rating condition. In follow-up analyses, we found that each pairwise comparison across prime type levels was statistically significant within the prime influence (all $ps < 0.037$) and, crucially, the unpleasant target influence (all $ps < 0.001$) and pleasant target influence (all $ps < 0.001$) conditions.

Discussion

In this experiment, judgments of both prime influence and target influence tracked the strength of primes. When primes were weakly valenced, participants were unlikely to report either prime or target influence; when primes were strongly valenced, participants were likely to report both prime and target influence. As such, this experiment clearly demonstrates that awareness of prime influence (as found in the prime influence rating condition) is compatible with (mistaken) awareness of target influence (as found in the target influence rating condition). These results are difficult to reconcile with Hughes et al.'s (2023) claim according to which "misattribution by definition cannot occur with awareness" (p. 1580). By contrast, these results are easily reconcilable with extensive prior evidence for the AMP effect emerging from an unintentional misattribution mechanism (Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2005, 2012) and with the possibility, discussed in Experiments 5–6, that judgments of prime influence emerge from a (potentially error-prone) inferential mechanism.

General discussion

Payne et al. (2005) introduced the affect misattribution procedure (AMP) as a measure of implicit evaluations, i.e., evaluations revealed under relatively automatic conditions. Specifically, these authors argued that the AMP indexes unintentionally expressed evaluative knowledge, and evidence in favor of this idea has been provided both in the original paper and in over a dozen follow-up studies (e.g., Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2012). As such, at this time, the construct validity of the AMP as a measure of unintentional evaluations is well supported by empirical evidence. Accordingly, the accuracy of the inferences drawn about substantive phenomena—such as impression formation, intergroup bias, and psychopathology—using this measure is not in any serious doubt, especially under a common (but not universally accepted) definition of implicit as unintentional (see, e.g., Devine, 1989; Ferguson & Cone, 2021; Melnikoff & Kurdi, 2022; Shiffrin & Schneider, 1977).

Notably, Payne et al. (2005) left open the question of whether, in addition to being unintentional, AMP effects also emerge beyond awareness of the influence of the primes on evaluative responses. As pointed out by several theoretical and methodological approaches to automaticity, the features of (un)intentionality and (un)awareness are not necessarily aligned with each other (e.g., Bargh, 1994; Melnikoff & Bargh, 2018; Moors, 2016; Moors & De Houwer, 2006). To reiterate the example cited multiple times throughout this paper, someone on a diet could feel an unwelcome and irresistible pull toward a delicious éclair displayed in a patisserie window, but they may or may not be aware of this pull. As such, the present results (as well as the results reported by Hughes et al., 2023) are uninformative with respect to whether the AMP measures unintentional evaluations but are potentially instructive with respect to another dimension of automaticity: awareness.

As explained above, we believe that the question of awareness deserves empirical scrutiny. Some classic treatments of implicit social cognition tend to focus on the issue of unawareness (to the exclusion of other automaticity criteria such as unintentionality) as the defining feature of implicit evaluations. For example, according to Greenwald and Banaji (1995) implicit evaluations are "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8). Participants being introspectively aware of their own implicit evaluations, as measured by the AMP, is obviously inconsistent with this definition. Similarly, in a recent theoretical piece, Gawronski et al. (2022) have highlighted the importance of studying unconscious evaluative processes. As such, whether the AMP measures evaluations in the absence of awareness (or in the absence of an intention to evaluate but with awareness of doing so) has important theoretical and practical implications.

In fact, based on results obtained by Hahn and colleagues (e.g., Hahn & Gawronski, 2019; Hahn et al., 2014; Rivers & Hahn, 2019) demonstrating correlations between predicted and actual implicit evaluations, Gawronski et al. (2022) have argued that currently available indirect measures, including the AMP, are not well equipped to address a central question of implicit social cognition research, namely whether social category information can influence downstream social behavior—including decisions in areas such as hiring and promotion, criminal sentencing, housing, and policing—in the absence of awareness. Morris and Kurdi (2023) conducted direct follow-up experiments to the work by Hahn and colleagues and showed that (*a*) participants are particularly accurate at predicting implicit evaluations that are highly correlated with explicit evaluations; (*b*) third-party observers are able to predict implicit evaluations based on demographic information alone; and (*c*) explicit evaluations account for considerably larger portions of the variance in predictions of implicit evaluations than implicit evaluations do. These findings call into question whether, and to what extent, the association between predicted and actual implicit evaluations is, in fact, mediated by privileged (and accurate) introspective awareness.

The results reported here are conceptually similar to those obtained by Morris and Kurdi (2023) in that they underscore that a statistical relationship between two variables (here, trial-by-trial self-reports of prime influence and the AMP effect), no matter how large, is insufficient to establish the direction of any potential causal relationship between them, as illustrated by the well-known example involving a high correlation between drowning deaths and ice cream consumption. Most germane to the present investigation, Hughes et al. (2023) found that the AMP effect was larger on trials in which participants report that their response had been influenced by the primes than on trials in which they did not. Based on this result, Hughes and colleagues called into question the validity of the AMP as a measure of implicit evaluations.

As we argued in the introduction, this conclusion is problematic on conceptual grounds alone: Whether AMP effects emerge beyond awareness or not does not directly speak to the measure's construct validity as an index of unintentional evaluations—the dimension of automaticity customarily taken to define the implicitness of the AMP. However, given the considerations outlined above, it is nonetheless theoretically instructive to probe whether, and to what degree, AMP effects are subject to accurate introspective access. For example, relevant results can be informative with respect

to the ability of the AMP to tap unconscious evaluations (Gawronski et al., 2022) and with respect to the nature of the misattribution mechanism assumed to give rise to the AMP effect (Payne et al., 2005). To summarize our conclusions, it seems unwarranted to assume that self-reported influence of the primes is a (pure) indication of privileged introspective processes. Moreover, at this time, no evidence is available to suggest that self-reported influence of the primes (whatever process such self-reports may emerge from) is a moderator of AMP effects, although the two are robustly correlated with each other both in past work by Hughes et al. (2023) and in the present experiments.

In Experiments 1–4 we found major fluctuations in the magnitude of the awareness effect and its relationship with the AMP effect. In Experiments 1–3, the proportion of influence-aware trials varied considerably depending on seemingly irrelevant factors, such as whether being influence-aware or being influence-unaware was the default option (or no default option was provided). Moreover, in Experiment 2, the size of the awareness effect was robustly correlated with participants' self-reported tendency to skip trials in order to get through the experiment as quickly as possible, thus suggesting that the prime influence measure may have been contaminated by construct-irrelevant factors. Finally, and crucially, in Experiment 4 we found that the relationship between awareness and the AMP effect was strongly moderated by prime valence such that aware and unaware AMP effects deviated especially strongly for extremely negative primes and were virtually identical for moderately valenced and positive (including extremely positive) primes.

Overall, these findings suggest that the influence awareness measure introduced by Hughes et al. (2023), at least if taken as an unadulterated expression of privileged introspective processes, is not as valid or as reliable as would be desirable. We believe that if a self-report item is to be the basis of claims about introspective awareness of AMP effects in the future, more resources will have to be devoted to ensuring adequate construct validity and measurement properties of this item (e.g., Flake & Fried, 2020; Hussey & Hughes, 2019). Moreover, even assuming adequate validity and reliability, the size of the awareness effect is unclear given that the proportion of aware trials fluctuated from 32% in Experiment 1 to 57% in the reversed condition of Experiment 2, with intermittent values obtained in Experiments 3–4.

However, despite these limitations, an awareness effect emerged consistently across Experiments 1–4. As such, in Experiments 5–7, we investigated the source of the awareness effect and the direction of the causal relationship between awareness effects and AMP effects. Based on classic (Bem, 1972; Nisbett & Wilson, 1977) and contemporary (Moutoussis et al., 2014) perspectives, we reasoned that the relationship might be the result of an inferential mechanism—namely, that participants may be more likely to report having been influenced if the prime and their response were evaluatively congruent than when they were not. In line with this possibility, we found an association between AMP effects and awareness effects even when (*a*) awareness judgments were made for third parties (Experiment 5) and (*b*) primes were presented post hoc (Experiment 6). Moreover, in Experiment 7, we demonstrated that this lean notion of awareness—a conscious inference that is available to self-report but does not presuppose privileged introspective access to one's mind—is consistent with an implicit misattribution mechanism. Specifically, in this experiment, prime strength moderated both prime influence and target influence judgments.

The latter finding warrants some additional discussion given assumptions in past work that awareness of prime influence may be incompatible with an unintentional misattribution mechanism (Hughes et al., 2023; Payne et al., 2005) and that, in this particular instance, two otherwise not necessarily aligned features of automaticity (unintentionality and unawareness) may mutually inform each other. For instance, Payne et al. (2005) note that "[a] great deal of research has shown that when people are aware of a potentially biasing influence, they often adjust their judgment to eliminate or even reverse the bias" (p. 279). However, this analysis does not appropriately consider a critical aspect of the AMP procedure, namely that whether a particular evaluative response was caused by the prime, the target, or a combination of both, is opaque to participants. Indeed, participants do not know whether they would have evaluated the target similarly or differently in the absence of the prime. As such, at least as long as this evaluative ambiguity persists (Oikawa et al., 2011; Ruys et al., 2012), participants may report being influenced by the prime although such influence is partly (or even fully) unintentional.

Taken together, the present findings provide direct evidence for (*a*) awareness effects emerging from a joint consideration of prime valence and response valence and (*b*) such awareness effects being caused by AMP effects rather than vice versa. Although the present findings provide direct evidence for this alternative causal mechanism, they notably do not rule out the causal mechanism originally proposed by Hughes et al. (2023), under which AMP effects are caused by awareness effects, also playing a role. However, the data by Hughes et al. (2023) provide no unique evidence for this causal direction, whereas the present data provide unique evidence for the opposite one.

As such, if Hughes and colleagues wish to make causal claims about awareness effects giving rise to AMP effects, the ball is presently in their court: Causal claims require experimental evidence, which was not provided in their paper or elsewhere. We believe that it is conceivable that the causal direction favored by Hughes and colleagues can operate some of the time and under some conditions, but for

now the totality of available evidence can be accounted for by the inferential mechanism proposed here (and for which, unlike for the opposite causal direction, we have provided indirect evidence in Experiments 2–4 and direct evidence in Experiments 5–7). In addition, we note again that even more well-supported claims about AMP effects being subject to introspective access (and perhaps even being caused by it) are not directly relevant to the issue of whether AMP performance reveals unintentional processes in social evaluation given that intentionality and awareness are conceptually and empirically distinct dimensions of automaticity.

In fact, although the present data do not conclusively rule out the causal mechanism conjectured by Hughes et al. (2023), several results are difficult to reconcile with this causal mechanism. First, if the influence measure is a direct reflection of participants' (accurate and privileged) introspective access to the contents of their minds, then why did influence awareness rates fluctuate to the degree that they did in Experiments 1–3 as a result of minor procedural variations? Second, if the influence awareness measure taps (mostly) introspective awareness, then why is it so robustly correlated with influence judgments made for third parties (Experiment 5)? Third, why did we observe a stronger (rather than weaker) relationship between AMP effects and influence awareness for third-party than for first-person judgments in Experiment 5? If anything, this finding seems to suggest that any privileged first-person information that participants possess about themselves may interfere with, rather than promote, influence awareness.

Although the present findings are informative regarding the automaticity conditions of the AMP, and specifically regarding the question of whether AMP effects can emerge in the absence of conscious awareness of prime influence, they leave at least three questions open, which we hope will be taken up in more detail in future work. The first question concerns the applicability of the present findings to the AMP as usually implemented in relevant research; the second question concerns the precise nature of the purported inferences made by participants when responding to the prime influence item; and the third question concerns the valence asymmetry effect obtained both by Hughes et al. (2023) and in all present experiments. Below we address each of these issues in turn.

First, to what extent can the present results be expected to generalize to AMPs as commonly used in relevant research? Indeed, the AMPs used in the present experiments deviate from AMPs used more customarily in that (*a*) they relied on strongly valenced primes (Experiments 1–3, 5, and 7) or no presentation of primes at all (Experiment 6) and (*b*) participants were asked to judge prime influence following each trial. We believe that the question of generalizability does not directly apply to Experiments 5–7 given that these experiments were designed to provide proof-of-concept demonstrations of the possibility that the pattern of means obtained by Hughes et al. (2023) can emerge from particular (inferential) processes. The question of generalizability emerges more straightforwardly with regard to Experiments 1–3, which were designed to document the existence and robustness of the awareness effect in the first place. In this context, Hughes et al. (2023) have already provided evidence for generalizability across the standard and modified AMPs both in the same (Experiment 3) and across different attitude domains (Experiment 4).

Given that we were able to replicate the findings of Experiment 2 from Hughes et al. (2023), we are optimistic that the findings of Experiments 3 and 4 would also be replicable. However, one important difference between the standard and the modified AMP procedure should be emphasized with respect to the size of the awareness effect. Specifically, Experiments 4–5 of Hughes et al. (2023) provided initial evidence that the statistical relationship between self-reports of prime influence and the AMP effect is larger for primes of extreme valence (IAPS images) than for primes of more moderate valence (images of Barack Obama and Donald Trump). We confirmed this result more directly in the present Experiment 4 by manipulating prime valence in a continuous manner. As such, given that they used exclusively extremely valenced primes in most of their experiments, Hughes et al. (2023) likely overestimated the relationship between awareness effects and AMP effects in a manner that does not generalize to the more mildly valenced primes used in the overwhelming majority of AMP research. In other words, no matter what causal mechanism is assumed to be underlying awareness effects, it seems that worries about the generality of a strong statistical relationship between self-reports of prime influence and responding on the AMP are overstated.

Second, what is the precise nature of the inferences made by participants when they are asked to respond to the measure of prime influence? In broad strokes, we believe that participants may be relying on the idea that evaluative congruence between a prime and a response (e.g., positive prime–pleasant response) is indicative of prime influence. The present Experiments 5 and 6 offer quite direct experimental evidence for this possibility given that those experiments provided participants with little (if any) other information on which to base their judgments of prime influence given that they were making those judgments either for third parties (Experiment 5) or in the absence of any actual primes having been presented (Experiment 6).

Asking participants to report the way in which they arrived at the relevant judgments may seem to present an even more direct way to provide support for this possibility. However, it is a truism in social cognition research that self-reports of internal processes can seldom be taken at face value (Nisbett & Wilson, 1977). Indeed, in past work

by Payne et al. (2012), participants agreed with the possibility of intentional prime influence when that possibility was presented to them as part of the question and with the possibility of unintentional prime influence when that possibility was presented to them as part of the question. Similarly, in the present experiments, participants' responding to the prime influence measure depended strongly on the default option (Experiment 2); moreover, participants readily acquiesced to the possibility that they indicated influence awareness to end trials more quickly. As such, perceived experimenter expectations and demand or reactance effects may also well have been operating across the present experiments as well as in past work by Hughes et al. (2023).

From the perspective of the inferential account proposed here, such effects are not surprising. Participants are faced with the virtually impossibly difficult task of making judgments about their own mental processes under conditions of extreme uncertainty. In fact, not even the experimenter is privy to the ground truth of whether a participant was influenced by the prime on a given trial, because such knowledge would presuppose access to the counterfactual (namely, how the participant would have responded in the absence of a prime). As such, participants are likely to rely on any information that they have at their disposal to respond to the challenging prime influence item. Future research could address in more detail the types of information that they use and the processes by which these judgments emerge. However, for the present purposes, it seems sufficient to demonstrate that inferential mechanisms (whatever their exact nature) can operate in the modified AMP procedure; and, indeed, the present results provide ample evidence for this possibility.

Third, why did a valence asymmetry effect emerge both in the studies conducted by Hughes et al. (2023) and in the present experiments? Specifically, why was the awareness effect stronger for negative primes (especially extremely negative primes) and weaker (or even nonexistent) for extremely positive primes? We can only speculate here, and we hope that future research will more directly explore the possibilities that we sketch out here.

One possibility is that this result is due to a negative potency bias and the relative risks involved in failing to detect negative compared with positive stimuli (Rozin & Royzman, 2001): Given this bias, participants may be more likely to spontaneously attend to negative primes than to positive ones, especially those that are extreme in valence (see also Brosch & Sharma, 2005; Öhman et al., 2001). And, once a stimulus is attended to, it is subsequently more likely to become part of an inferential process (e.g., an explanation for an AMP response) than if it is not. Alternatively, or in addition, when negative stimuli are present, they can interfere with the operation of propositional processes (Kurdi et al., 2022b). Specifically, in this case, propositional processes that may have assigned causal credit for the participant's negative behavioral response to a source other than the prime (e.g., the target stimulus or an internal state unrelated to either) may have operated to a lesser degree in the presence of extremely negative primes than they would have otherwise. We hope that future work will more directly examine these and other potential mechanisms.

The valence asymmetry effect was not expected a priori; however, critically, it does not invalidate any of our previous conclusions. In fact, we note that the valence asymmetry effect emerged consistently across experiments involving first-person judgments (Experiments 1–4), third-party judgments (Experiment 5), and even in a design that did not feature the presentation of valenced primes at all (Experiment 6). This convergence suggests that whatever processes are implicated in giving rise to the valence asymmetry effect generalize across these instantiations of the procedure. This result, in turn, provides additional evidence for the idea that judgments of prime influence may emerge from similar mechanisms whether they are made for the self (Experiments 1–4), for others (Experiment 5), or in situations where no prime influence is possible in the first place (Experiment 6).

Conclusion

To conclude, our methodological recommendation to those looking to use an indirect measure to capture unintentional evaluations is that a large body of evidence, accumulated over two decades, now suggests that the AMP is a valid measure of such evaluations (Bar-Anan & Nosek, 2016; Gawronski & Ye, 2013, 2015; Mann et al., 2019; Payne et al., 2005, 2012). The AMP was not designed to measure evaluations in the absence of awareness, and those in need of a paradigm capturing this type of evaluation are best served using procedures that were specifically developed for this purpose, including subliminal priming (e.g., Greenwald et al., 1995). At the same time, these procedures are known for their relatively noisy and parameter-dependent nature, and therefore require ample resources and expertise to implement. It is conceivable that a variant of the AMP will be able to serve as a viable alternative going forward; however, given that the possibility of awareness contributing to AMP effects has not been conclusively eliminated, we do not recommend such use at the present time.

AMP effects have been shown to occur unintentionally, which makes the AMP a measure of implicit evaluations under a wide range of definitions of automaticity. Here we investigated whether AMP effects also emerge in the absence of awareness and found no evidence suggesting that they do not. Although a statistical relationship of variable size exists between AMP effects and self-reports of prime influence, these self-reports seem to reflect an inferential mechanism rather than privileged introspective access to the contents

of one's mind. Moreover, these inferences appear to be a result, rather than a cause, of AMP effects. Although the present data do not rule out the possibility of reverse causation, at present, the idea that AMP effects emerge from privileged first-person awareness of prime influence remains mere conjecture.

Data Availability All data (including trial-level AMP data) are available via the Open Science Framework (https://osf.io/wfksp/).

References

Bar-Anan, Y., & Nosek, B. A. (2011). Reporting intentional rating of the primes predicts priming effects in the Affective Misattribution Procedure. *Personality and Social Psychology Bulletin, 38*(9), 1194–1208. https://doi.org/10.1177/0146167212446835

Bar-Anan, Y. & Nosek, B. A. (2016). *Misattribution of claims: Comment on Payne et al., 2013*. PsyArXiv. https://doi.org/10.31234/osf.io/r75xb

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). Guilford Press.

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models*. ArXiv. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Elsevier. https://doi.org/10.1016/s0065-2601(08)60024-6

Brosch, T., & Sharma, D. (2005). The role of fear-relevant stimuli in visual search: A comparison of phylogenetic and ontogenetic stimuli. *Emotion, 5*(3), 360–364. https://doi.org/10.1037/1528-3542.5.3.360

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

Cooley, E., & Payne, B. K. (2016). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin, 43*(1), 46–59. https://doi.org/10.1177/0146167216675331

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

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(e33400), 1507–1517. https://doi.org/10.3389/fpsyg.2014.00781

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition, 25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

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

Ferguson, M. J., & Cone, J. (2021). The role of intentionality in priming. *Psychological Inquiry, 32*(1), 38–40. https://doi.org/10.1080/1047840x.2021.1889839

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*(2), 203–210. https://doi.org/10.1037/h0041593

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30*(6), 652–668. https://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry, 33*(3), 139–155. https://doi.org/10.1080/1047840x.2022.2106750

Gawronski, B., & Ye, Y. (2013). What drives priming effects in the Affect Misattribution Procedure? *Personality and Social Psychology Bulletin, 40*(1), 3–15. https://doi.org/10.1177/0146167213502548

Gawronski, B., & Ye, Y. (2015). Prevention of intention invention in the Affect Misattribution Procedure. *Social Psychological and Personality Science, 6*(1), 101–108. https://doi.org/10.1177/1948550614543029

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

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology, 124*(1), 22–42. https://doi.org/10.1037/0096-3445.124.1.22

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y. & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 U.S. presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 241–253. https://doi.org/10.1111/j.1530-2415.2009.01195.x

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology, 116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hughes, S., Cummins, J., & Hussey, I. (2023). Effects on the Affect Misattribution Procedure are strongly moderated by influence awareness. *Behavior Research Methods, 55*(4), 1558–1586. https://doi.org/10.3758/s13428-022-01879-4

Hussey, I. & Cummins, J. (2022). *Evidence against effects on the Affect Misattribution Procedure being unaware: AMP effects involve construct-irrelevant individual differences*. PsyArXiv. https://psyarxiv.com/8k94v

Hussey, I., & Hughes, S. (2019). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy, 3*(2), 159–186. https://doi.org/10.1017/bpp.2018.43

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70*(1), 172–194. https://doi.org/10.1037/0022-3514.70.1.172

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science, 302*(5469), 1338–1339. https://doi.org/10.1126/science.1091721

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*(1), 59–99. https://doi.org/10.1023/a:1005296023424

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697–720. https://doi.org/10.1037/0003-066x.58.9.697

Katz, J. H., Mann, T. C., Shen, X., Goncalo, J. A., & Ferguson, M. J. (2022). Implicit impressions of creative people: Creativity evaluation in a stigmatized domain. *Organizational Behavior and Human Decision Processes, 169*, 104116. https://doi.org/10.1016/j.obhdp.2021.104116

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kurdi, B., Hussey, I., Stahl, C., Hughes, S., Unkelbach, C., Ferguson, M. J. & Corneille, O. (2022a). Unaware attitude formation in the surveillance task? Revisiting the findings of Moran et al. (2021). *International Review of Social Psychology, 35*(1). https://doi.org/10.5334/irsp.546

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2022b). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology, 124*(6), 1174–1202. https://doi.org/10.1037/pspa0000329

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods, 49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Lang, P. J., Bradley, M. M. & Cuthbert, B. N. (2008). *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. Technical report A-8*. University of Florida, Gainesville.

Lee, K. M., Lindquist, K. A., & Payne, B. K. (2018). Constructing bias: Conceptualization breaks the link between implicit bias and fear of Black Americans. *Emotion, 18*(6), 855–871. https://doi.org/10.1037/emo0000347

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the Affect Misattribution Procedure. *Journal of Personality and Social Psychology, 116*(3), 349–374. https://doi.org/10.1037/pspa0000146

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry, 33*(3), 185–192. https://doi.org/10.1080/1047840x.2022.2106759

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology, 67*(1), 263–287. https://doi.org/10.1146/annurev-psych-122414-033550

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Morris, A. & Kurdi, B. (2023). Awareness of implicit attitudes: Large-scale investigations of mechanism and scope. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0001464

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition, 25*(100), 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology, 64*(5), 723–739. https://doi.org/10.1037/0022-3514.64.5.723

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295x.84.3.231

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Oikawa, M., Aarts, H., & Oikawa, H. (2011). There is a fire burning in my heart: The role of causal attribution in affect transfer. *Cognition & Emotion, 25*(1), 156–163. https://doi.org/10.1080/02699931003680061

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466–478. https://doi.org/10.1037/0096-3445.130.3.466

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2012). Intention invention and the Affect Misattribution Procedure. *Personality and Social Psychology Bulletin, 39*(3), 375–386. https://doi.org/10.1177/0146167212475225

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology, 46*(2), 367–374. https://doi.org/10.1016/j.jesp.2009.11.001

Payne, B. K., & Lundberg, K. (2014). The Affect Misattribution Procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Perszyk, D. R., Lei, R. F., Bodenhausen, G. V., Richeson, J. A., & Waxman, S. R. (2019). Bias at the intersection of race and gender: Evidence from preschool-aged children. *Developmental Science, 22*(3), e12788. https://doi.org/10.1111/desc.12788

Rivers, A. M., & Hahn, A. (2019). What cognitive mechanisms do people reflect on when they predict IAT scores? *Personality and Social Psychology Bulletin, 45*(6), 878–892. https://doi.org/10.1177/0146167218799307

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/s15327957pspr0504_2

Ruys, K. I., Aarts, H., Papies, E. K., Oikawa, M., & Oikawa, H. (2012). Perceiving an exclusive cause of affect prevents misattribution. *Consciousness and Cognition, 21*(2), 1009–1015. https://doi.org/10.1016/j.concog.2012.03.002

Schreiber, F., Neng, J. M. B., Heimlich, C., Witthöft, M., & Weck, F. (2014). Implicit affective evaluation bias in hypochondriasis: Findings from the Affect Misattribution Procedure. *Journal of Anxiety Disorders, 28*(7), 671–678. https://doi.org/10.1016/j.janxdis.2014.07.004

Shiffrin, R. M. & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295x.84.2.127

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

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tucker, R. P., Wingate, L. R., Burkley, M., & Wells, T. T. (2018). Implicit association with suicide as measured by the Suicide Affect Misattribution Procedure (S-AMP) predicts suicide ideation. *Suicide and Life-Threatening Behavior, 48*(6), 720–731. https://doi.org/10.1111/sltb.12392

Wentura, D. (2000). Dissociative affective and associative priming effects in the lexical decision task: Yes versus no responses to word targets reveal evaluative judgment tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 456–469. https://doi.org/10.1037/0278-7393.26.2.456

Wentura, D., Müller, P., & Rothermund, K. (2014). Attentional capture by evaluative stimuli: Gain- and loss-connoting colors boost the additional-singleton effect. *Psychonomic Bulletin & Review, 21*(3), 701–707. https://doi.org/10.3758/s13423-013-0531-z

Williams, A., & Steele, J. R. (2017). Examining children's implicit racial attitudes using exemplar and category-based measures. *Child Development, 21*(1), 55–17. https://doi.org/10.1111/cdev.12991

Wood, S.N. (2017). *Generalized additive models: An introduction with R* (2nd edition). Chapman and Hall/CRC.