Journal of Experimental Social Psychology 88 (2020) 103950

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Attitudes as prepared reflexes☆

David E. Melnikoff*, Robert Lambert, John A. Bargh

Yale University, USA

ARTICLE INFO

Keywords:
Attitudes
Prepared reflexes
Goals
Implicit evaluation
Action control

ABSTRACT

When people plan to respond to a stimulus S with an action R, they hold an S-R association in working memory. Such S-R associations are called prepared reflexes. In the present investigation, we explored the possibility that prepared reflexes play a central role in evaluative processing. Specifically, we tested the hypothesis that attitudes toward a given stimulus S (*i*) become more positive when prepared reflexes associate S with a positively valenced action representation R+, and (*ii*) become more negative when prepared reflexes associate S with a negatively valenced action representation R-. We found support for this hypothesis across 6 studies while ruling out alternative mechanisms including cognitive dissonance, self-perception, approach-avoid training, and biased scanning. We conclude by discussing the implications of our findings for the predictive validity of implicit attitude measures.

1. Introduction

Attitudes and actions emerge from deeply interconnected cognitive systems. Each receives input from the other, allowing attitudes to shape actions, and actions to shape attitudes (Lavender and Hommel, 2007; Neumann, Förster, & Strack, 2003). Though this linkage is well-known, some of its implications, we suggest, are underappreciated. For attitudes research, it means that investigators can leverage theories of action control to formulate novel hypotheses about attitude formation and change. This approach, however, is rarely taken (cf. Eder & Klauer, 2007, 2009; Van Dessel, Eder, & Hughes, 2018). Principles of action control are seldom exported to the attitudes literature, perhaps concealing some important sources of attitude change. We propose that one such source is a cognitive structure called the *prepared reflex*, a structure that is central to theories of action control, but absent from theories of attitudes. If we are right, the prepared reflex would constitute a new mechanism of attitude change, one that may help resolve some long-standing puzzles concerning both the nature and predictive power of human likes and dislikes. In what follows, we define prepared reflexes and describe how they might shape attitudes.

2. Attitudes as prepared reflexes

The concept of the prepared reflex was introduced over a century ago by Exner (1879), and later refined by Woodworth (1938). Here is how it works. When people plan to respond to a stimulus S with an action R, they hold an S-R association in working memory (Cole, Braver, & Meiran, 2017; Exner, 1879; Hommel, 2000; Meiran, Cole, & Braver, 2012; Woodworth, 1938). For as long as that S-R association is in working memory, the mere perception of S will reflexively activate a mental representation of R.[1] For instance, if a person plans to execute a left-hand keypress whenever they see the letter D, that person will hold in working memory an association between D and < left-hand keypress > , allowing the perception of D to reflexively activate < left-hand keypress > . Such S-R associations are called prepared reflexes (Hommel, 2000) – "prepared" in the sense that they are formed intentionally (i.e. by planning actions), and "reflexes" in the sense that they are activated unintentionally (i.e., by external stimuli). As Woodworth (1938) put it: "The reaction is involuntary, i.e., no new will impulse is needed after the entrance of the stimulus in order that the reaction shall follow. The only voluntary act is the preparation" (p. 305). This notion has been revived by modern theories of action control (e.g., Gollwitzer, 1999; Hommel, 2000), and conclusive evidence has accumulated in support of it (e.g., Cohen, Bayer, Jaudas, & Gollwitzer, 2008; Cohen-Kdoshay & Meiran, 2009; Cole, Braver, & Meiran, 2017; Meiran et al., 2012; Miles & Proctor, 2008).

Critically, prepared reflexes need not be associations between stimulus representations and affectively *neutral* action representations

☆ This paper has been recommended for acceptance by Marlone Henderson.

* Corresponding author at: Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06511, USA.
E-mail address: david.melnikoff@yale.edu (D.E. Melnikoff).

[1] This results in a predisposition to perform R, which may or may not be translated into actual behavior. This translation is influenced by numerous factors, including the complexity of the represented action, the presence or absence of inhibitory signals, and decision thresholds for action execution.

https://doi.org/10.1016/j.jesp.2019.103950
Received 2 March 2019; Received in revised form 13 December 2019; Accepted 30 December 2019

such as < left-hand keypress > — they may be associations between stimulus representations and *valenced* action representations such as < help > and < harm > (Eder & Klauer, 2009). Indeed, action representations come to include the affective valence with which they typically co-occur (Barrett, 2017; Eder & Hommel, 2013; Lavender & Hommel, 2007). The action representation < help > , for instance, is positive (because helping typically co-occurs with positive affect; (Batson, 1987; Haidt, 2003; Trivers, 1971)) and the action representation < harm > is negative (because harming typically co-occurs with negative affect; (Haidt, 2003; Rai, Valdesolo, & Graham, 2017)).

We propose that by linking stimulus representations to affectively valenced action representations, prepared reflexes change attitudes. Specifically, when a person plans to respond to a stimulus S with a valenced action R, a prepared reflex will emerge that activates R upon exposures to S. If R has a positive valence (+), then activation will spread to +, causing evaluative responses toward S to become more positive. Conversely, if R has a negative valence (−), then activation will spread to -, causing evaluative responses toward S to become more negative.

Four specific hypotheses flow from our theory of how prepared reflexes shape attitudes. We present each hypothesis below, and then show how these hypotheses distinguish prepared reflexes from related mechanisms of attitude change.

2.1. The inaction hypothesis

Every prepared reflex has an underlying action plan (i.e. a plan to respond to a stimulus S with an action R). According to the Inaction Hypothesis, such action plans need never be executed for prepared reflexes to change attitudes. Suppose, for instance, that a defense attorney plans to help her client by presenting exonerating evidence, resulting in a prepared reflex that activates < help > (a positive action representation) on perception of the client. According to the Inaction Hypothesis, the attorney would evaluate her client more positively without ever attempting the helpful action — merely *planning* to help her client is sufficient to induce the more positive attitude. The Inaction Hypothesis follows straightforwardly from the premise that prepared reflexes emerge from action *plans* rather than actions.

2.2. The transience hypothesis

The Transience Hypothesis states that prepared reflexes change attitudes only for as long their underlying action plans are in place — once action plans are terminated, their effects on attitudes are eliminated. Returning to our attorney, the Transience Hypothesis says that her prepared reflex would cause her to evaluate her client more positively only for as long she plans to perform the helpful action. If the attorney abandons her plan — because the case is dismissed, for instance — then her original, relatively negative attitude toward her client will be reinstated. The Transience Hypothesis follows from two premises: (*i*) prepared reflexes are held in working memory, and (*ii*) working memory has limited capacity, which people occupy only when necessary (Miller, 1956; Simon, 1982). Prepared reflexes are thus unloaded from working memory once their underlying action plans are terminated, and thus no longer relevant, so as not to waste processing capacity (Fagot, 1994; Meiran, 2000, 2005; Meiran et al., 2012). In other words, S-R links are unloaded from working memory once an individual no longer plans to perform R toward S, at which point activation has no means of spreading from S to R, thereby eliminating any attitude change.

2.3. The additivity hypothesis

Attitude change is subject to constraints, some of which are imposed by the features of the stimulus being evaluated. One such constraint is ambiguity. Attitudes toward unambiguously positive or negative stimuli tend to be harder to change than attitudes toward ambiguous stimuli (Fazio, 2007; Gregg, Seibt, & Banaji, 2006; Kunda & Thagard, 1996; Pyszczynski & Greenberg, 1987). It is harder, for instance, to change attitudes toward cancerous tumors than chocolate cake. Chocolate cake can be evaluated positively by attending to its positive feature (e.g., taste), or negatively by attending to its negative features (e.g., unhealthiness). This strategy is unlikely to work for cancerous tumors, which lack positive features.

Another constraint on attitude change is the preexisting valence of the target stimulus. Attitudes toward negative stimuli are harder to change than attitudes toward positive stimuli. This negativity bias animates the dictum “Bad is stronger than good,” which applies to a broad range of psychological phenomena (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Mende-Siedlecki, Baron, & Todorov, 2013; Rozin & Royzman, 2001).

Although ambiguity and valence typically constrain attitude change, the Additivity Hypothesis stipulates that neither factor constrains the evaluative effects of prepared reflexes. A given prepared reflex should have the same sized effect on attitudes irrespective of the target stimulus' ambiguity and preexisting valence. The Additivity Hypothesis follows from the premise that prepared reflexes change attitudes via spreading of activation from a stimulus representation S to the valence of an action representation R, effectively adding the valence of R to S. If the evaluative effects of prepared reflexes are indeed additive, then their effects are, by definition, unmoderated by the preexisting contents of the stimulus representation.

2.4. The dissociation hypothesis

The Dissociation Hypothesis states that prepared reflexes directly alter *implicit* (i.e. unintentionally activated) attitudes, but not *explicit* (i.e. intentionally reported) attitudes. This hypothesis follows from two premises. First, whereas explicit evaluation entails affirming beliefs about the valence of stimuli (e.g., “S is positive”), implicit evaluation does not — implicit evaluation merely entails the unintentional activation of valence on stimulus perception (e.g., thinking “positive” unintentionally on perception of S) (De Houwer, 2014; Gawronski & Bodenhausen, 2006; Olson & Fazio, 2004).

The second premise underlying the Dissociation Hypothesis is that prepared reflexes are implemented by S-R associations, which are not beliefs about the valence of S, but can activate valence unintentionally on perception of S (i.e., S-R associations are not beliefs of the form “S is positive,” but can cause people to think “positive” unintentionally on perception of S). From these two premises flows the Dissociation Hypothesis: prepared reflexes directly alter implicit attitudes, but not explicit attitudes, because prepared reflexes activate valence unintentionally on stimulus perception, but are not beliefs about the valence of stimuli.

Although prepared reflexes should not change explicit attitudes directly, they may do so indirectly via separate processes outside the action control system. One such processes is misattribution. When a prepared reflex activates the valence of an action representation, that valence may get misattributed to the stimulus with which the action was associated. Our attorney, for instance, may infer that she thinks “positive” on perception of her client not because her planned action is positive, but because her client is. This misattribution would alter the attorney's explicit attitude toward her client. Critically, misattribution can be distinguished from prepared reflexes: Unlike prepared reflexes, misattribution violates the Additivity Hypothesis. People are likelier to misattribute positive feelings to positive (relative to negative) stimuli, and to misattribute negative feelings to negative (relative to positive) stimuli (Taylor & Fiske, 1978).

Self-perception (Bem, 1972) is another process, outside of the action control system, through which prepared reflexes may indirectly alter explicit attitudes. Specifically, people may infer that a stimulus is good

or bad on the basis of having planned to respond to that stimulus positively or negatively (e.g., "I planned to perform a positive action towards S… I only plan to perform positive actions toward good S… therefore S is good") (see Van Dessel, Hughes, & De Houwer, 2018).

Self-perception can be distinguished from prepared reflexes on two counts. First, unlike prepared reflexes, self-perception violates the Transience Hypothesis. Returning to our attorney, suppose she infers that her client is good on the basis of having planned a helpful action (e.g., "I planned to help my client… I only plan to help good people… therefore my client is good"). This inference ("My client is good") would persist if the trial were dismissed, since both premises ("I planned to help my client" and "I only plan to help good people") would remain valid (Van Dessel et al., 2018). Accordingly, the attorney's new explicit attitude would violate the Transience Hypothesis.

The attorney new explicit attitude would violate the Additivity Hypothesis as well. The more negative a stimulus is, the less likely people are to infer that they like that stimulus on the basis of planning a positive action. Reflecting on her plan to help, the attorney is less likely to infer that she likes her client if her client is a mass murderer than if her client is a tax evader. The same principle applies in reverse: The more positive a stimulus is, the less likely people are to infer that they dislike that stimulus on the basis of planning a negative action (see Centerbar & Clore, 2006).

In summary, the Dissociation Hypothesis says that prepared reflexes directly alter implicit attitudes, but not explicit attitudes. Their effects on explicit attitudes are indirect, mediated by processes outside the action control system. Thus, prepared reflexes should be associated with different patterns of implicit and explicit attitude change: the former, but not the latter, should adhere to both the Transience Hypothesis and the Additivity Hypothesis.

3. Alternative mechanisms of goal-driven attitude change

We have theorized that prepared reflexes play a fundamental role in shaping attitudes. In the context of the broader literature, this theory describes a novel effect of goals on evaluative processing. Indeed, we have said that the formation of action plans — a core component of goal pursuit — changes attitudes via prepared reflexes. It is therefore important to distinguish prepared reflexes from previously established mechanisms by which goals influence evaluative processing, especially those that allow action plans to shape attitudes. Cognitive dissonance, self-perception, approach-avoid (AA) training, and biased scanning (Olson & Stone, 2005) deserve special attention, as each involves a causal path from action plans to attitude change. In what follows, we briefly describe these mechanisms, and distinguish them from prepared reflexes by indicating the above hypotheses (Inaction, Transience, Dissociation, and/or Additivity) with which they are incompatible (see Table 1).

Table 1
Prepared reflexes and alternative mechanisms of attitude change.

Mechanisms	Hypotheses			
	Inaction	Transience	Additivity	Dissociation
Prepared reflexes	✓	✓	✓	✓
Cognitive dissonance	✓	–	–	–
Self-perception	✓	–	–	–
AA training	–	–	✓	✓
Biased scanning	✓	✓	–	–

Note. A check mark indicates that the process may be compatible the corresponding hypothesis, whereas a dash indicates that the process is incompatible with the corresponding hypothesis.

3.1. Cognitive dissonance

According to dissonance theory (Festinger, 1957), setting or pursuing goals that are inconsistent with one's attitudes may arouse psychological discomfort, which one may resolve via attitude change. For instance, a self-identifying introvert might experience cognitive dissonance after setting a goal to welcome a new neighbor, and might resolve that dissonance by forming more positive attitudes toward that neighbor.

Cognitive dissonance is incompatible with the Dissociation Hypothesis, the Transience Hypothesis, and the Additivity Hypothesis. Regarding the Dissociation Hypothesis, hundreds of studies have shown that cognitive dissonance changes explicit attitudes, but none have shown that cognitive dissonance changes implicit attitudes (Olson & Stone, 2005). Prominent theories even stipulate that implicit attitudes are immune to cognitive dissonance (Gawronski & Bodenhausen, 2006, 2011; Gawronski & Brannon, 2018; Gawronski & Strack, 2004; McConnell & Rydell, 2014; Rydell & McConnell, 2006) — a prediction for which there exists direct support (Gawronski & Strack, 2004). Accordingly, we are aware of no empirical or theoretical reasons to suspect that cognitive dissonance directly changes implicit attitudes, but not explicit attitudes.

Regarding the Transience Hypothesis, cognitive dissonance creates stable attitudes in long-term memory, and thus their effects persist after the dissonance-arousing action plan has been unloaded from working memory (Lieberman, Ochsner, Gilbert, & Schacter, 2001). Dissonance-induced attitude change has been shown to persist for more than several weeks (Collins & Hoyt, 1972; Festinger & Carlsmith, 1959; Freedman, 1965; Sénémeaud & Somat, 2009) and to occur even in individuals with no explicit memory for their dissonance-arousing behavior (Lieberman et al., 2001).

Regarding the Additivity Hypothesis, dissonance-reduction processes are fundamentally constrained by the plausibility of the resulting attitude, and thus are constrained by stimulus features (Elliot & Devine, 1994; Festinger, 1957; Olson & Stone, 2005; Pyszczynski & Greenberg, 1987). If it is illogical to conclude that a stimulus is positive or negative on the basis of one's goal (or pursuit thereof), then that goal is unlikely to change attitudes via cognitive dissonance. This means that cognitive dissonance has relatively weak effects on attitudes toward unambiguous and/or negative stimuli (e.g., murderers), and has relatively strong effects on attitudes toward ambiguous and/or positive stimuli (e.g., new neighbors) (Elliot & Devine, 1994; Festinger, 1957; Olson & Stone, 2005).

3.2. Self-perception

According to self-perception theory (Bem, 1972), people may infer that they like or dislike stimuli after observing themselves set or pursue goals that involve acting positively or negatively toward those stimuli, respectively. For instance, having observed oneself agree to welcome a new neighbor, one might infer that they like that neighbor.

Like cognitive dissonance, self-perception is incompatible with the Dissociation Hypothesis, the Transience Hypothesis, and the Additivity Hypothesis. Indeed, self-perception is similar to cognitive dissonance in that (*i*) there are no empirical or theoretical reasons to suspect that self-perception directly changes implicit attitudes but not explicit attitudes, (*ii*) self-perception creates stable (explicit) attitudes in long-term memory (J. M. Olson & Stone, 2005), and (*iii*) self-perception is subject to a plausibility constraint (Bem, 1972; Fazio, 1987).

3.3. Approach-avoid training

One method of changing attitudes is to have people repeatedly approach one stimulus and to avoid another. This procedure is called approach-avoid (AA) training, and it tends to produce a preference for the approached stimulus over the avoided stimulus (Jones, Vilensky,

Vasey, & Fazio, 2013; Kawakami, Phills, Steele, & Dovidio, 2007; Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011; Woud, Maas, Becker, & Rinck, 2013). Researchers have argued that AA training changes attitudes by creating associations in long-term memory between a target stimulus and < approach > (a positively valenced action representation) or < avoid > (a negatively valenced action representation); but see (Van Dessel, Eder, and Hughes, 2018). Unlike prepared reflexes, AA training is incompatible with the Inaction Hypothesis and the Transience Hypothesis.

AA training is incompatible with the Inaction Hypothesis because AA training involves actual pairings between actions and stimuli – by definition, effects of AA training are dependent on overt action. Regarding the Transience Hypothesis, AA training changes evaluative associations in long-term memory, and thus its effects on attitudes persist after the relevant action plans have been unloaded from working memory (Eberl et al., 2013).

3.4. Biased scanning

Goals can change attitudes by directing attention toward positive or negative features of the target stimulus, or by reconstruing features of a target stimulus in a positive or negative light (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017; Fazio, 2007; Ferguson & Bargh, 2004; Fujita, 2011; Melnikoff & Bailey, 2018; Sinclair & Kunda, 1999; Wittenbrink, Judd, & Park, 2001). This process – called biased scanning (which encompasses attention modulation and cognitive change/reconstrual) – causes attitudes to change in ways that make action plans easier to implement (Berkman et al., 2017; Ferguson & Bargh, 2004; Fujita, 2011). For instance, someone with a dieting goal may evaluate a slice of cake more negatively, and thus resist eating it, by shifting her attention toward the cake's high fat content and away from its delicious taste.

Biased scanning is incompatible with the Dissociation Hypothesis and the Additivity Hypothesis. Regarding the Dissociation Hypothesis, a large body work suggests that biased scanning induces equivalent patterns of implicit and explicit attitude change (Berkman et al., 2017; Brendl & Higgins, 1996; Cabanac, 1971; Fishbach, Shah, & Kruglanski, 2004; Fitzsimons & Fishbach, 2010; Fitzsimons & Shah, 2008; Fujita & Carnevale, 2012; (Fujita, Trope, Liberman, & Maya, 2006); Kunda, 1990; Melnikoff & Bailey, 2018; J. M. Olson & Stone, 2005; (Orehek & Forest, 2016) Pyszczynski & Greenberg, 1987(Sinclair, Lowery, Hardin, & Colangelo, 2005). In contrast, we are aware of no evidence, empirical or theoretical, that biased scanning alters implicit attitudes directly, and explicit attitudes only indirectly.

Regarding the Additivity Hypothesis, biased scanning is constrained by the contents of the stimulus being evaluated (Pyszczynski & Greenberg, 1987). One constraint is ambiguity — biased scanning cannot “work” unless a stimulus has both positive and negative contents for which to scan. For instance, if all one knows about someone is that they are a murderer, a scan for positive content would likely fail, leaving attitudes toward the murderer unchanged. A second constraint is valence — attention is attracted to negative stimuli more than positive stimuli (Baumeister et al., 2001; Ito, Larsen, Smith, & Cacioppo, 1998; Pratto & John, 1991; Rozin & Royzman, 2001). Attitudes toward negative stimuli, therefore, are less responsive to biased scanning than are attitudes toward positive stimuli (Baumeister et al., 2001; Cacioppo, Gardner, & Berntson, 1997; Cone & Ferguson, 2015; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). The relationship between biased scanning and stimulus features is thus interactive rather than additive.

4. The present studies

We ran 6 studies testing the hypothesis that attitudes emerge from prepared reflexes, while systematically ruling out each of the alternative mechanisms reviewed above. One of these studies replicated the findings of Study 1 using a nearly identical procedure. This study is included in the meta-analysis, but omitted from the main text (details are available in Supplemental materials).

Across all 6 studies, our basic approach was to have participants plan an action toward a target person, and to manipulate whether that action was positive or negative. Plans to perform a positive action R should create prepared reflexes that activate R's positivity upon exposures to the target person S, resulting in more positive attitudes toward S. Conversely, plans to perform a negative action R should create prepared reflexes that activate R's negativity upon exposures to the target person S, resulting in more negative attitudes toward S. Showing that attitudes change as a function of valenced action plans, however, is not sufficient to support our claim that attitudes change as a function of prepared reflexes. What we must show is that valenced action plans change attitudes in a manner consistent with the Inaction Hypothesis, the Transience Hypothesis, the Dissociation Hypothesis, and the Additivity Hypothesis (see Table 1). We did so as follows.

First, we ensured that participants never implemented their action plans. Consequently, effects of valenced action plans on attitudes would be consistent with the Inaction Hypothesis, as they could not be attributed to actual pairings between actions and stimuli. Second, we manipulated whether participants unloaded their prepared reflexes from working memory prior to attitude assessment. All participants formed an action plan, but we told some participants right before measuring their attitudes that they would not be performing their action plan after all. No longer expecting to execute their action plan, these participants should unload the prepared reflex that corresponds to their action plan from working memory – this, according to the Transience Hypothesis, should completely eliminate any attitude change. Third, we ensured that the target person's valence was unambiguously positive or negative.[2] For instance, one target person risked his own life to save someone from drowning, another target person murdered his friend in cold blood, and another target person was Adolf Hitler. The Additivity Hypothesis stipulates that attitudes toward these target people change as a function of valenced action plans, despite the fact that their valence is unambiguous, and irrespective of whether their valence is positive or negative.

Finally, we tested the Dissociation Hypothesis by measuring both implicit and explicit attitudes. In each individual study, we support the Dissociation Hypothesis indirectly by documenting significant effects on type of attitude and non-significant effects on the other. We provide direct support in the meta-analysis that follows Study 5 by documenting significant differences between implicit and explicit attitude change.

4.1. Measuring attitudes

We used a self-report measure to assess participants' explicit attitudes, and the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) to assess implicit attitudes. The AMP, which is one of the most widely used measures of automatic evaluation, has been shown to be reliable and predictive of actual behavior (see Payne & Lundberg, 2014).

Although Bar-Anan and Nosek (2012) raised concerns that participants who take the AMP may intentionally rate the primes (calling into question the “implicitness” of the AMP), more recent work by Payne et al. (2013) disconfirmed this possibility. Specifically, Payne et al. (2013) found that the findings of Bar-Anan and Nosek (2012) are attributable to participants making post-hoc confabulations to explain their responses, and that attitudes as measured by the AMP indeed reflect unintentional evaluations of the prime stimuli (i.e. implicit attitudes). Moreover, if our participants *were* to intentionally rate the

[2] Ambiguity is a continuous dimension, but for stylistic reasons we describe stimuli as being “ambiguously” or “unambiguously” positive or negative. By “ambiguous” we mean “relatively ambiguous” and by “unambiguous” we mean “relatively unambiguous”.

primes, this would cause AMP performance to align with explicit attitudes, which in turn would disconfirm the Dissociation Hypothesis. Accordingly, intentional responding to the primes is (*i*) highly unlikely in light of the available evidence Payne et al. (2013), and (*ii*) would actually *decrease* our likelihood of supporting one of our key predictions.

4.2. Determining sample size

The most complex predicted effect was a three-way interaction between two between-subjects factors and one within-subjects factor, all with two levels. An a priori power analysis – which assumed a modest correlation between repeated measures ($r = 0.30$) and a small-to-medium effect size ($\eta_p^2 = 0.02$) – revealed that 192 participants who were required to have an 80% finding such an effect. Thus, in each study, we aimed for a final sample of at least 200 participants. To further maximize both statistical power and the reliability of key effect sizes, we conducted a meta-analysis of all 6 studies, the results of which appear after Study 5. Finally, we note that all measures, manipulations, and exclusions are disclosed in all studies, and data were always collected in a single wave.

5. Study 1

5.1. Participants

Using Amazon Mechanical Turk (MTurk), we recruited 338 native English speakers living in the United States. All participants were recruited in a single wave, and were asked to summarize key procedural details throughout this and every other study to maximize attention and comprehension. We chose all exclusion criteria a priori: participants who failed to correctly indicate whether they would help or harm the target, had mean response times under 200 ms on either of the two AMPs (indicative of random responding), or reported recognizing the Chinese pictographs were eliminated ($N = 62$). This resulted in a final sample size of 276 (62% Female), ages 19 to 71 ($M = 39$, 95% CI [38, 40]). The exclusion rate (18%) did not differ by condition. We report all exclusions, manipulations, and measures. Participant gender did not moderate any of the results reported in this or any other study.

5.2. Procedure

We introduced participants to the study by explaining that we are interested in people's ability to ignore goal-relevant imagery. To explore this topic, our cover story explained, participants would play a game called Attorney at Law, and complete visual processing tasks designed to measure their ability to ignore game-relevant images. Participants then proceeded to learn about Attorney at Law.

First, participants learned about the main character: a man named Francis West. Participants saw a photo of Francis (a middle-aged White male) and read a story about him. All participants read that Francis went to the beach with his friend Roger, who got pulled under water by a strong current. What happened next in the story constituted our Stimulus Content manipulation, which had two conditions: "unambiguously positive" and "unambiguously negative." In the unambiguously positive condition, Francis acted heroically: he risked his own life to save Roger, but despite doing everything he could, the current was too strong and Roger died. In the unambiguously negative condition, Francis acted malevolently: Roger was able to get his head above water, but Francis forced Roger under, drowning him in cold blood. Participants in both conditions read that Francis was accused of murder and will stand trial (this accusation was true in the unambiguously negative condition, and false in the unambiguously positive condition). After reading one of the two stories, participants completed the first of two AMPs, allowing us to establish baseline implicit attitudes toward Francis as both levels of Stimulus Content.

Next, participants were randomly assigned to the role of prosecuting attorney or defense attorney in Francis' trial. This constituted our manipulation of Plan Valence, which had two conditions: "positive action plan" and "negative action plan." Participants in the negative action plan condition learned that they would *harm* Francis – a type of action that typically co-occurs with negative affect. These participants learned that they would be the prosecuting attorney, and thus would present the jury with "as much negative, incriminating evidence as possible." Participants in the positive action plan condition learned that they would *help* Francis – a type of action that typically co-occurs with positive affect. These participants learned that they would be the defense attorney, and thus would present the jury with "as much positive, exonerating evidence as possible." If prepared reflexes shape attitudes, then participants in the positive action plan condition should evaluate Francis more positively due to an association in working memory between Francis and a positively valenced action representation; conversely, participants in the negative action plan condition should evaluate Francis more negatively due to an association in working memory between Francis and a negatively valenced action representation. Furthermore, the Additivity Hypothesis stipulates that these effects be unmoderated by Stimulus Content.

To ensure that participants were motivated, we stated that $10 would be rewarded to the 10 best attorneys (i.e. those who present the most piece of incriminating or exonerating evidence). In fact, participants never actually played Attorney at Law (allowing us to test the Inaction Hypothesis), so we randomly selected 10 participants to receive the reward. At this point, we asked all participants to indicate whether their job was to prosecute or defend Francis West — we decided a priori that participants who answered this question incorrectly would be excluded from subsequent analyses.

Next, we explained the rules of Attorney at Law. We told all participants that images of Francis plus various distractor images would appear on their screen during the trial. Prosecuting attorneys learned that Francis would appear on their screen with "INNOCENT" under his photo — pressing their space bar within 2 s of seeing Francis would remove "INNOCENT," and disprove a piece of exonerating evidence. Conversely, defense attorneys learned that Francis would appear on their screen with "GUILTY" under his photo — pressing their space bar within 2 s of seeing Francis would remove "GUILTY," and disprove a piece of incriminating evidence. To ensure that all participants formed the appropriate action plan, we had participants type the following plan three times: "If I see Francis West, then I will eliminate the word ["INNOCENT"/"GUILTY"] as fast as possible!"

Next came the Unload manipulation, which allowed us to test the Transience Hypothesis. The Unload manipulation had two conditions: "load" (in which prepared reflexes were loaded in working memory throughout attitude assessment) and "unload" (in which prepared reflexes were unloaded from working memory prior to attitude assessment). Participants in the unload condition read a message stating that would not play Attorney at Law after all, as it is not compatible with their operating system; they would finish the visual processing tasks and skip to the end of the survey (we further explained that, of the participants who do not play Attorney at Law, we would randomly select 10 to receive the $10 bonus). After reading this message, participants in the unload condition completed the second AMP. Participants in the unload condition thus completed the second AMP without prepared reflexes in working memory linking Francis to valenced action representations — the corresponding action plans had been terminated. In contrast, participants in the load condition received the same message as those in the unload condition, but not until *after* completing the second AMP (and explicit attitude measures; see below). Thus, among participants in the load condition, prepared reflexes corresponding to valenced action plans were loaded in working memory throughout attitude assessment. The Transience Hypothesis stipulates that attitudes toward Francis change among participants in the load condition but not among participants in the unload condition, even though all

participants went through the same process of forming a valenced action plan.

After completing the second AMP, participants reported their explicit attitudes toward Francis. Participants indicated on 7-point scales how much they like or dislike Francis, how positively or negatively they feel toward Francis, and how good or bad Francis is (Cronbach's $\alpha = 0.95$). We measured explicit attitudes in order to assess the Dissociation Hypothesis, which stipulates that implicit attitudes, but not explicit attitudes, adhere to the Additivity Hypothesis and the Transience Hypothesis.

Next, participants in the load condition received the message that participants in the unload condition had received previously, indicating that they would not have enough time to complete the trial. Participants then completed a demographics survey and were debriefed. Note that participant never performed their planned action toward Francis. Any effects of Plan Valence on attitudes are thus consistent with the Inaction Hypothesis.

5.3. Affect misattribution procedure

The AMP (Payne et al., 2005) consisted of 10 practice trials and 60 critical trials. Only the critical trials were analyzed. Of the 60 critical trials, half were target trials and half were control trials. Target trials began with the photograph of Francis that participants had seen earlier in the study, and control trials began with a photograph of an unfamiliar, middle-aged white male randomly selected from a set of 5. After 75 ms., the photograph was immediately replaced with a randomly selected Chinese pictograph. The pictograph was displayed for 100 ms, then immediately followed by a backward mask (random black and white noise). The backward mask remained on the screen until participants made a response.

Participants' task was to indicate whether the Chinese pictograph was more or less visually pleasant than average. Following Payne et al. (2005), we warned that the preceding photographs can influence their judgments of the Chinese pictographs, and that we are interested in their ability to ignore such influence and provide an unbiased judgment of each pictograph. Previous research has shown that people misattribute their automatic evaluation of the initial image to their feelings about the pictograph, thus providing a measure of people's spontaneous and unintentional evaluations of the photographs of target vs. control faces.

We computed automatic evaluations of Francis as follows. First, for each participant, we determined the proportion of Chinese pictographs rated as pleasant separately for target trials and control trials. Second, we subtracted the proportion of *pleasant* responses on control trials from the proportion of *pleasant* responses on target trials. Thus, positive values reflect positive automatic evaluations of Francis relative to control, and negative values reflect negative automatic evaluations of Francis relative to control. By computing difference scores between target and control trials we controlled for response bias toward liking or disliking White male faces in general, and/or finding Chinese pictographs to be visually pleasant or unpleasant overall.

5.4. Results

5.4.1. Implicit attitudes

We conducted a mixed analysis of variance (ANOVA) with Plan Valence, Unload, Stimulus Content, and Time predicting implicit attitudes toward Francis. The main effect of Stimulus Content was significant, $F(1, 268) = 12.06$, $p = .001$, $\eta_p^2 = 0.043$, such that participants implicitly liked unambiguously positive Francis ($M = 3.73$, 95% CI [0.26, 7.2]) more than unambiguously negative Francis ($M = -4.62$, 95% CI [−7.83, −1.4]). This effect was qualified by a Stimulus Content x Time interaction, $F(1, 268) = 7.28$, $p = .007$, $\eta_p^2 = 0.026$, such that the effect of Stimulus Content was stronger after plan formation ($M_{\text{UnambiguouslyPositive}} = 6.3$, 95% CI [1.51, 11.08]; $M_{\text{UnambiguouslyNegative}} = -6.41$, 95% CI [−10.84, −1.97]), $F(1, 268) = 14.69$, $p < .001$, $\eta_p^2 = 0.052$, versus before ($M_{\text{UnambiguouslyPositive}} = 1.16$, 95% CI [−2.31, 4.63]; $M_{\text{UnambiguouslyNegative}} = -2.83$, 95% CI [−6.04, 0.39]), $F(1, 268) = 2.75$, $p = .099$, $\eta_p^2 = 0.01$.

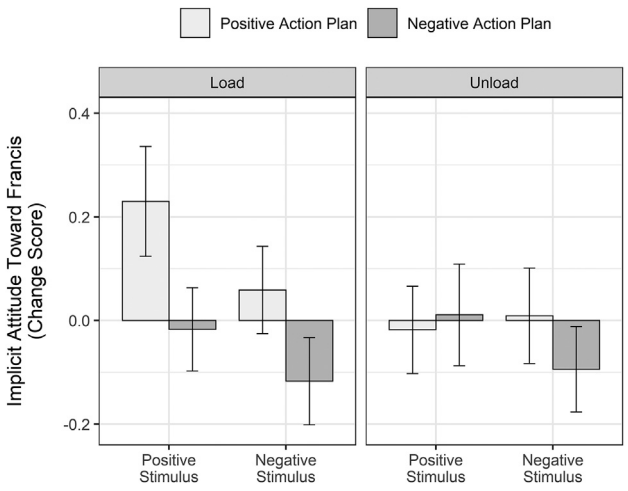

Fig. 1. Change in implicit attitudes toward Francis (after goal induction - before goal induction) as a function of Plan Valence (positive vs. negative), Stimulus Content (unambiguously positive vs. unambiguously negative), and Unload (load vs. unload). Error bars represent 95% CIs.

The main effect of Plan Valence was significant, $F(1, 268) = 10.77$, $p = .001$, $\eta_p^2 = 0.039$, such that implicit attitudes toward Francis were more positive among participants who formed a positive action plan ($M = 3.5$, 95% CI [0.05, 6.95]) versus a negative action plan ($M = -4.39$, 95% CI [−7.62, −1.15]). This effect of Plan Valence was qualified by Time, $F(1, 268) = 14.78$, $p < .001$, $\eta_p^2 = 0.052$, and Unload, $F(1, 268) = 5.07$, $p = .025$, $\eta_p^2 = 0.019$. However, both of these two-way interactions were further qualified by a Plan Valence x Unload x Time interaction, $F(1, 268) = 7.26$, $p = .007$, $\eta_p^2 = 0.026$ (Fig. 1).

We decomposed the three-way interaction by evaluating the Plan Valence x Time interaction separately at both levels of Unload. Consistent with the Transience Hypothesis, the Plan Valence x Time interaction was significant in the load condition, $F(1, 268) = 21.51$, $p < .001$, $\eta_p^2 = 0.12$, but not in the unload condition, $F(1, 268) = 0.66$, $p = .418$, $\eta_p^2 = 0.006$.

Next, among participants in the load condition, we assessed the effect of Plan Valence before and after plan formation. This effect was significant after plan formation, $F(1, 268) = 26.13$, $p < 001$, $\eta_p^2 = 0.089$, but not before, $F(1, 268) = 0.65$, $p = .422$, $\eta_p^2 = 0.002$, due to increased positivity in the positive action plan condition, and decreased positivity in the negative action plan condition. Specifically, participants who formed a positive action plan implicitly liked Francis more after plan formation ($M = 15.1$, 95% CI [8.14, 22.07]) versus before ($M = 0.67$, 95% CI [−4.38, 5.71]), $F(1, 268) = 17.54$, $p < .001$, $\eta_p^2 = 0.061$, and participants who formed a negative action plan implicitly liked Francis less after plan formation ($M = -8.76$; 95% CI [−14.77, −2.75]) versus before ($M = -2.05$; 95% CI [−6.41, 2.3]), $F(1, 268) = 5.08$, $p = .025$, $\eta_p^2 = 0.019$. Consistent with the Additivity Hypothesis, Stimulus Content did not moderate the Plan Valence x Unload x Time interaction, $F(1, 268) = 2.46$, $p = .118$, $\eta_p^2 = 0.009$. No other effects were significant.

5.4.2. Explicit attitudes

We conducted a between-subjects ANOVA to test the effects of Plan Valence, Unload, and Stimulus Content on explicit attitudes toward Francis ($\alpha = 0.95$). The main effect of Stimulus Content was significant, such that participants explicitly liked unambiguously positive Francis

(M = 5.11, 95% CI [4.89, 5.33]) more than unambiguously negative Francis (M = 2.91, 95% CI [2.71, 3.11]), F(1, 268) = 214.49, p < .001, η_p^2 = 0.445. The main effect of Plan Valence was significant as well: explicit attitudes toward Francis were more positive among participants who formed a positive action plan (M = 4.5, 95% CI [4.28, 4.71]) versus a negative action plan (M = 3.52, 95% CI [3.32, 3.73]), F(1, 268) = 41.88, p < .001, η_p^2 = 0.135.

Unexpectedly, the Plan Valence x Stimulus Content interaction was not significant, F(1, 268) = 2.73, p = .099, η_p^2 = 0.01. Plan Valence had similar effects on explicit attitudes toward unambiguously positive Francis ($M_{Positive}$ = 5.47, 95% CI [5.16, 5.79]; $M_{Negative}$ = 4.75, 95% CI [4.45, 5.04]); F(1, 268) = 10.79, p = .001, η_p^2 = 0.039), and unambiguously negative Francis ($M_{Positive}$ = 3.52, 95% CI [3.23, 3.81]; $M_{Negative}$ = 2.3, 95% CI [2.02, 2.57]); F(1, 268) = 35.73, p < .001, η_p^2 = 0.118). We had expected Stimulus Content to moderate the effect of Plan Valence on explicit (but not implicit) attitudes. It is possible that our manipulation of Stimulus Content was not strong enough to induce this dissociation — a possibility we explore in Study 3.

Nevertheless, we found support for the Dissociation Hypothesis in the form of a nonsignificant Plan Valence x Unload interaction, F(1, 268) = 0.929, p = .336, η_p^2 = 0.003. Unlike implicit attitudes, explicit attitude were responsive to Plan Valence in both the load condition ($M_{Positive}$ = 4.6, 95% CI [4.29, 4.92]; $M_{Negative}$ = 3.49, 95% CI [3.21, 3.76]); F(1, 268) = 27.87, p < .001, η_p^2 = 0.094), and the unload condition ($M_{Positive}$ = 4.39, 95% CI [4.09, 4.68]; $M_{Negative}$ = 3.56, 95% CI [3.26, 3.86]); F(1, 268) = 15.05, p < .001, η_p^2 = 0.053). In other words, explicit attitudes, but not implicit attitudes, violated the Transience Hypothesis.

5.5. Discussion

The results of Study 1 (and its replication; see Supplemental materials) provide support for the hypothesis that prepared reflexes shape attitudes: valenced action plans changed attitudes in line with the Inaction, Transience, Additivity, and Dissociation hypotheses. Consistent with the Inaction Hypothesis, valenced action plans that were never performed changed attitudes. Consistent with the Transience Hypothesis, valenced action plans changed (implicit) attitudes only for as long as they were in working memory. Consistent with the Dissociation Hypothesis, prepared reflexes appeared to directly affect implicit attitudes, but not explicit attitudes (i.e. the former, but not the latter, adhered to the Transience Hypothesis). Consistent with the Additivity Hypothesis, attitudes changed despite the fact that the target person's valence was unambiguous, and irrespective of whether the target person's valence was positive or negative. This constellation of findings is uniquely consistent with the operation of prepared reflexes (see Table 1).

6. Study 2

We designed Study 2 to replicate Study 1 while exploring potential boundary conditions. The procedure was identical to that of Study 1, with two exceptions. First, we introduced a Plan Specificity manipulation. Research has shown that plans with a specific, if-then format (e.g., "*If* I see Francis West, *then* I will remove the word 'INNOCENT' as fast as possible!") often have greater self-regulatory benefits than unspecific plans (e.g., "I will do my best to disprove exonerating evidence about Francis West!") (see Gollwitzer, 1999). Accordingly, the results of Study 1 may overestimate the strength and generalizability of the underlying phenomenon due to the fact that all participants in these studies formulated their action plans in a specific, if-then format. To test this possibility, we manipulated Plan Specificity by assigning each participant to one of two conditions: "specific plan" or "unspecific plan." In the specific plan condition, participants typed "If I see Francis West, then I will remove the word ["GUILTY"/"INNOCENT"] as fast as possible!" three times, as in Study 1. In the unspecific plan condition, participants typed "I will do my best to disprove [incriminating/exonerating] evidence about Francis West!" three times.

Besides manipulating Plan Specificity, we removed the Unload manipulation in order to limit the complexity of the experimental design. All participants received the message indicating that their operating system is not compatible with Attorney at Law after they completed the second AMP and explicit attitude measures. Every other aspect of Study 2 was identical to Study 1.

6.1. Participants

Using MTurk, we recruited 271 native English speakers living in the United States to complete Study 2. All data were collected in a single wave. Data from 35 participants was excluded using the same exclusion criteria as in the previous studies, resulting in a final sample size of 236 (58% Female), ages 18 to 71 (M = 35, 95% CI [34, 36]). The exclusion rate (13%) did not differ by condition.

6.2. Results

6.2.1. Implicit attitudes

We ran a mixed ANOVA with Plan Valence, Plan Specificity, Stimulus Content, and Time predicting implicit attitudes toward Francis. The main effect of Stimulus Content was significant, F(1, 228) = 13.87, p < .001, η_p^2 = 0.057, such that participants implicitly liked unambiguously positive Francis (M = 4.99, 95% CI [1.93, 8.05]) more than unambiguously negative Francis (M = −3.27, 95% CI [−6.39, −0.15]). The main effect of Plan Valence was significant as well, F(1, 228) = 4.8, p = .029, η_p^2 = 0.021: implicit attitudes toward Francis were more positive among participants who formed a positive action plan (M = 3.29, 95% CI [0.21, 6.37]) versus a negative action plan (M = −1.57, 95% CI [−4.67, 1.53]). This effect was qualified by Time, F(1, 228) = 16.1, p < .001, η_p^2 = 0.066 Fig. 2.

Thus, we assessed the effect of Plan Valence before and after plan formation. As predicted, the effect of Plan Valence was significant after plan formation, F(1, 228) = 13.36, p < .001, η_p^2 = 0.055, but not before, F(1, 228) = 1.21, p = .273, η_p^2 = 0.005. This effect was driven by increased positivity among participants who formed positive action plans, and decreased positivity among participants who formed negative action plans. Among participants who formed a positive action plan, implicit attitudes toward Francis were more positive after plan

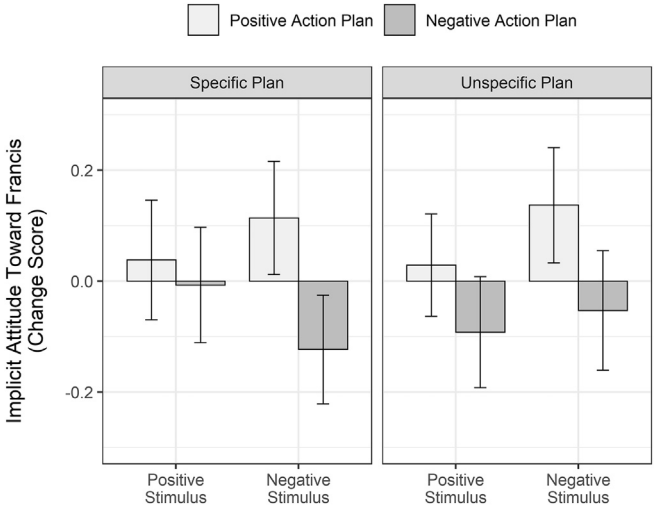

Fig. 2. Change in implicit attitudes toward Francis (after plan formation - before plan formation) as a function of Plan Valence (positive action plan vs. negative action plan), Stimulus Content (unambiguously positive vs. unambiguously negative) and Plan Specificity (specific action plan vs. unspecific action plan). Error bars represent 95% CIs.

formation (M = 7.26, 95% CI [2.6, 11.93]) versus before (M = −0.68, 95% CI [−3.91, 2.54]), F(1, 228) = 9.3, p = .003, η_p^2 = 0.039, and among participants who formed a negative action plan toward Francis, implicit attitudes toward Francis were more negative after plan formation (M = −5.01; 95% CI [−9.7, −0.32]) versus before (M = 1.87; 95% CI [−1.38, 5.11]), F(1, 228) = 6.90, p = .009, η_p^2 = 0.029.

Consistent with the Additivity Hypothesis, Stimulus Content did not moderate the Plan Valence x Time interaction, F(1, 228) = 3.08, p = .081, η_p^2 = 0.013. Besides the Stimulus Content x Plan Valence x Plan Specificity interaction, F(1, 228) = 4.13, p = .043, η_p^2 = 0.018 — which is not of theoretical interest given the absence of an effect of Time — no other effects were significant, including the Plan Valence x Plan Specificity x Time interaction, F(1, 228) = 0.04, p = .848, η_p^2 < 0.001.

6.2.2. Explicit attitudes

We conducted a between-subjects ANOVA to test the effects of Plan Valence, Plan Specificity, and Stimulus Content on explicit attitudes toward Francis (α = 0.96). The main effect of Stimulus Content was significant, F(1, 228) = 183.20, p < .001, η_p^2 = 0.446, such that participants explicitly liked unambiguously positive Francis (M = 5.14, 95% CI [4.92, 5.37]) more than unambiguously negative Francis (M = 2.89, 95% CI [2.66, 3.13]). The main effect of Plan Valence was significant as well, F(1, 228) = 22.94, p < .001, η_p^2 = 0.091: explicit attitudes toward Francis were more positive among participants who formed a positive action plan (M = 4.42, 95% CI [4.19, 4.65]) versus a negative action plan (M = 3.62, 95% CI [3.39, 3.85]).

As in Study 1 (and its replication; see Supplemental materials), the Plan Valence x Stimulus Content interaction was not significant, F(1, 228) = 0.51, p = .475, η_p^2 = 0.002. Plan Valence had a similar effects on explicit attitudes toward unambiguously positive Francis ($M_{Positive}$ = 5.6, 95% CI [5.28, 5.92]; $M_{Negative}$ = 4.69, 95% CI [4.36, 5.01]); F(1, 228) = 15.47, p < .001, η_p^2 = 0.064), and unambiguously negative Francis ($M_{Positive}$ = 3.23, 95% CI [2.9, 3.56]; $M_{Negative}$ = 2.56, 95% CI [2.23, 2.89]); F(1, 228) = 8.13, p = .005, η_p^2 = 0.034). No other effects were significant, including the Plan Valence x Plan Specificity interaction, F(1, 228) = 0.15, p = .701, η_p^2 = 0.001.

6.3. Summary

A direct manipulation of plan specificity had no effect on patterns of attitude change. Accordingly, the results of Study 2 both replicate and extend the results of Studies 1 by indicating that prepared reflexes shape attitudes even when their underlying action plans are relatively unspecific.

7. Study 3

We designed Study 3 to provide a stronger test of the Additivity Hypothesis. The procedure was identical to that of Study 1, with two exceptions. First, we replaced the name, photo, and description of Francis West with those of Adolf Hitler. Because Hitler is among the most negative people in history, this design choice permits an extremely conservative test of the Additivity Hypothesis. Indeed, according to the Additivity Hypothesis, the effects of action plans on attitudes should be just as large in Study 3 as they were in Studies 1–2, despite our having replaced a novel target person (Francis West) with the most notorious mass murderer in history.

The second way in which Study 3 differed from Study 1 is that it did not include a Stimulus Content manipulation. Our use of Hitler as the target stimulus made such a manipulation impossible.

7.1. Participants

Using MTurk, we recruited 301 native English speakers living in the United States. All participants were recruited in a single wave. We excluded participants according to the same criteria as in the previous studies (N = 50) resulting in a final sample size of 251 (61% Female), ages 18 to 67 (M = 35, 95% CI [34, 37]). The exclusion rate (17%) did not differ by condition.

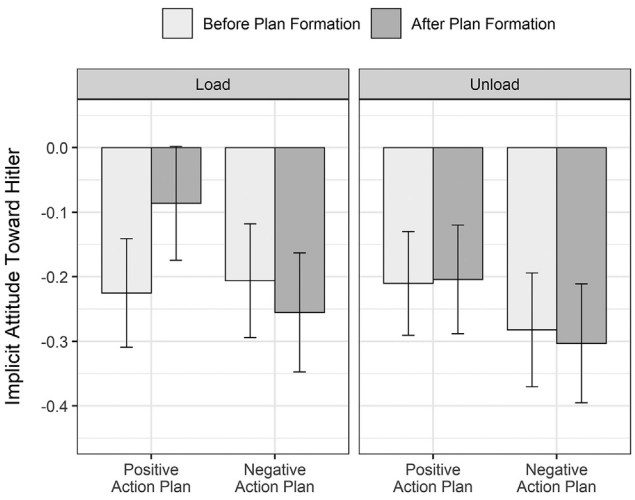

Fig. 3. Implicit attitudes toward Hitler as a function of Time (before plan formation vs. after plan formation), Plan Valence (positive vs. negative), and Unload (load vs. unload). Error bars represent 95% CIs.

7.2. Results

7.2.1. Implicit attitudes

We ran a mixed ANOVA with Plan Valence, Unload, and Time predicting implicit attitudes toward Hitler. We found a Plan Valence x Time interaction, F(1, 247) = 8.60, p = .004, η_p^2 = 0.034, which was qualified by the predicted Plan Valence x Time x Unload interaction, F(1, 247) = 4.85, p = .029, η_p^2 = 0.019 (Fig. 3).

Decomposing this three-way interaction, we found support for the Transience Hypothesis: the Plan Valence x Time interaction was significant in the load condition, F(1, 247) = 12.79, p < .001, η_p^2 = 0.078, but not the unload condition, F(1, 247) = 0.26, p = .611, η_p^2 = 0.003. Among participants in the load condition, we assessed the effect of Plan Valence before and after plan induction. This effect was significant after plan formation, F(1, 247) = 6.66, p = .01, η_p^2 = 0.026, but not before, F(1, 247) = 0.09, p = .76, η_p^2 < 0.001. This is because participants who formed a positive action plan evaluated Hitler more positively after plan formation (M = −8.59, 95% CI [−17.47, 0.28]) versus before (M = −22.53, 95% CI [−31.05, −13.99]), F(1, 247) = 14.89, p < .001, η_p^2 = 0.057. Among participants who formed a negative action plan, attitudes toward Hitler did not change over time, F(1, 247) = 1.64, p = .202, η_p^2 = 0.007. This null effect is likely is due to the fact that evaluations of Hitler were extremely negative at baseline; participants could not dislike Hitler any more than they already did. No other effects were significant.

7.2.2. Explicit attitudes

Unsurprisingly, explicit attitudes toward Hitler were extremely negative, with the mean falling significantly below the neutral midpoint of the scale (M = 1.39, 95% CI [1.27, 1.5]), t(250) = 44.54, p < .001, d = 2.81. To explore how these attitudes related to Plan Valence and Unload, we conducted a between-subjects ANOVA. Consistent with the Dissociation Hypothesis, we found that explicit attitudes — unlike implicit attitudes — were unaffected by Plan Valence, F(1, 247) = 0.52, p = .470, η_p^2 = 0.002: Explicit evaluations of Hitler were no more positive among participants who formed a positive action plan (M = 1.43, 95% CI [1.28, 1.59]) than participants who formed a

negative action plan (M = 1.35, 95% CI [1.18, 1.52]). This null effect was not moderated by Unload, $F(1, 247) = 3.32$, $p = .07$, $\eta_p^2 = 0.013$. Plan Valence had a null effect on explicit attitudes in both the load condition ($M_{PositivePlan}$ = 1.43, 95% CI [1.28, 1.59]; $M_{NegativePlan}$ = 1.33, 95% CI [1.09, 1.57]); $F(1, 247) = 3.16$, $p = .08$, $\eta_p^2 = 0.013$), and the unload condition ($M_{PositivePlan}$ = 1.24, 95% CI [1.02, 1.46]; $M_{NegativePlan}$ = 1.37, 95% CI [1.13, 1.61]); $F(1, 247) = 0.62$, $p = .433$, $\eta_p^2 = 0.002$).

7.3. Summary

Study 3 provides further support for the idea that prepared reflexes shape attitudes. The formation of action plans once again changed attitudes in a manner consistent with the Inaction, Dissociation, Transience, and Additivity hypotheses. Support for the Additivity Hypothesis was particularly strong: forming a positive action plan toward Hitler increased implicit liking of Hitler, despite Hitler's extreme and unambiguous negativity. Moreover, among participants in the load condition, the size of the effect of forming a positive action plan toward Hitler ($\eta_p^2 = 0.057$) was no smaller than the size of the effect of forming a positive action plan toward Francis across the first three studies, regardless of whether Francis was unambiguously positive ($\eta_p^2 = 0.026$) or unambiguously negative ($\eta_p^2 = 0.026$).

The results of Study 3 strongly support the Dissociation Hypothesis as well. Unlike implicit attitude change, which was unaffected by the switch from Francis to Hitler, explicit attitude change was completely eliminated; explicit attitudes, but not implicit attitudes, violated the Additivity Hypothesis. This marks the second dissociation between implicit and explicit attitude change, the first being the adherence of implicit but not explicit attitudes to the Transience Hypothesis (see Study 1).

8. Study 4

The purpose of Study 4 was to conceptually replicate our previous findings using a novel methodology. Specifically, we replaced the Attorney at Law game with a different scenario.

8.1. Participants

Using MTurk, we recruited 413 native English speakers living in the United States to complete Study 4. All participants were recruited in a single wave. We excluded participants according to the same criteria as in the previous studies (N = 42), resulting in a final sample size of 371 (64% Female), ages 18 to 68 (M = 33, 95% CI [32, 34]). The exclusion rate (10%) did not differ by condition.

8.2. Procedure

Introducing participants to the study, we explained that we are interested in the psychological experience of doctors who administer capital punishment. To help us, our cover story explained, participants were told that they would “take part” in the execution process by completing an extremely mild simulation that involves mixing the drugs used for lethal injections.

Next, participants learned about and saw a picture of Ken Morwitz. Ken's picture was a mugshot of a white, middle-aged male, different from Francis. To further assess the Additivity Hypothesis, we described Ken as extremely and unambiguously negative. Specifically, we told participants that Mr. Morwitz committed and confessed to the murders of two teenage girls and is scheduled to receive a lethal injection.

Next, participants completed the first AMP measuring their implicit attitudes toward Ken. The AMP we used in Study 4 was identical to those used in the previous studies apart from the fact that the critical prime was the photo of Ken Morwitz. After completing the first AMP, we told participants that, although medical personnel tasked with administering the lethal injection typically carry out their duties, they sometimes attempt to save the prisoner's life. To account for this in our study, we explained, participants would be randomly assigned to one of two conditions: some must try to terminate Ken Morwitz's life (a negatively valenced action plan) and some must try to save Ken Morwitz's life (a positively valenced action plan). We then randomly assigned participants to one of these two conditions, thereby manipulating Plan Valence.

Participants who planned to save Ken's life were told that those who make the 10 least effective mixtures would receive a $10 bonus; participants who planned to terminate Ken's life were told that those who make the 10 most effective mixtures would receive a $10 bonus. Additionally, we asked participants to indicate whether their goal was to save or terminate Ken's life. We decided a priori to exclude from analyses all participants who answered this question incorrectly. Also, to provide further evidence that specific if-then plans are not required for Plan Valence to shape attitudes, we had participants type the following, unspecific plan three times: “I will do my best to [terminate/save] Ken Morwitz's life”.

Next came the Unload manipulation, which was identical to the one we used in Studies 1 and 3, and which allowed us to test the Transience Hypothesis. After prepared reflexes were or were not unloaded, participants completed the second AMP. We then had participants report their explicit attitudes toward Ken using the same three items from Studies 1–3, which allows us to test the Dissociation Hypothesis. Participants then completed a demographics survey and were debriefed. As in every other study, no participants ever performed their action plan. Any effects of Plan Valence on attitudes are thus consistent with the Inaction Hypothesis.

8.3. Results

8.3.1. Implicit attitudes

We performed a mixed ANOVA to explore the effects of Plan Valence, Unload, and Time on implicit attitudes toward Ken. There was a main effect of Time, $F(1, 367) = 7.05$, $p = .008$, $\eta_p^2 = 0.019$, such that implicit attitudes toward Ken were more positive after plan formation (M = −13.93, 95% CI [−16.89, −10.37]) versus before (M = −17.43, 95% CI [−20.73, −14.12]), and a main effect of Unload, $F(1, 367) = 8.47$, $p = .004$, $\eta_p^2 = 0.023$, such that implicit attitudes toward Ken were more positive in the load condition (M = −11.14, 95% CI [−15.29, −6.89]) versus the unload condition (M = −19.92, 95% CI [−24.15, −15.68]). These main effects were qualified by a Plan Valence x Unload x Time interaction, $F(1, 367) = 3.95$, $p = .048$, $\eta_p^2 = 0.011$ (Fig. 4).

Decomposing the three-way interaction, we found support for the Transience Hypothesis: the Plan Valence x Time interaction was significant among participants in the load condition, $F(1, 367) = 4.17$, $p = .042$, $\eta_p^2 = 0.022$, but not among participants in the unload condition, $F(1, 367) = 0.61$, $p = .436$, $\eta_p^2 = 0.003$. Next, among participants in the load condition, we explored the effect of Plan Valence before and after plan formation. As predicted, the effect of Plan Valence was significant after plan formation, $F(1, 367) = 8.06$, $p = .005$, $\eta_p^2 = 0.021$, but not before, $F(1, 367) < 0.01$, $p = .990$, $\eta_p^2 < 0.001$. This is because participants who formed a positive action plan implicitly evaluated Ken more positively after plan formation (M = −3.41, 95% CI [−9.96, 3.15]) versus before (M = −11.56, 95% CI [−18.19, −4.93]), $F(1, 367) = 8.06$, $p = .005$, $\eta_p^2 = 0.021$. Among participants who formed a negative action plan, Time did not affect implicit attitudes toward Ken, $F(1, 367) < 0.001$, $p = .990$, $\eta_p^2 < 0.001$, likely due to a floor effect. No other effects were significant.

8.3.2. Explicit attitudes

Overall, explicit attitudes toward Ken were extremely negative, with the mean falling significantly below the neutral midpoint of the scale

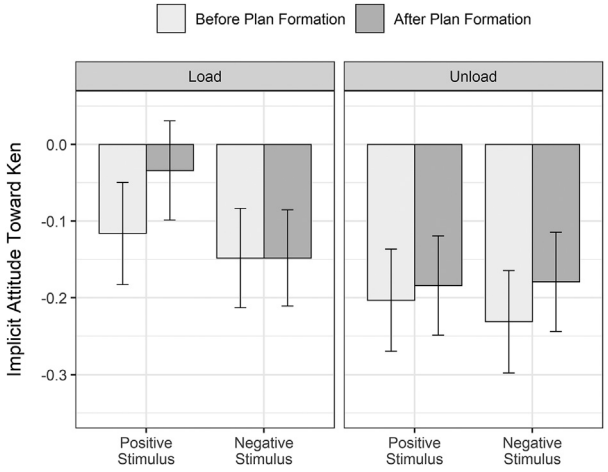

Fig. 4. Implicit attitudes toward Ken relative to control faces as a function of Time (before plan induction vs. after plan induction), Plan Valence (positive vs. negative), and Unload (load vs. unload). Error bars represent 95% CIs.

(M = 2.30, 95% CI [2.19, 2.41]), $t(370) = 30.51$, $p < .001$. To explore how these attitudes related to Plan Valence and Unload, we conducted a between-subjects ANOVA. The main effect of Plan Valence was significant, $F(1, 367) = 6.42$, $p = .012$, $\eta_p^2 = 0.017$, such that participants who formed a positive action plan (M = 2.44, 95% CI [2.29, 2.59]) explicitly evaluated Ken more positively than participants who formed a negative action plan (M = 2.16, 95% CI [2.01, 2.31]). The main effect of Unload was significant as well, $F(1, 367) = 6.1$, $p = .014$, $\eta_p^2 = 0.016$: participants in the load condition (M = 2.43, 95% CI [2.29, 2.59]) explicitly evaluated Ken more positively versus participants in the unload condition (M = 2.17, 95% CI [2.01, 2.32]).

Consistent with the Dissociation Hypothesis, the Plan Valence x Unload interaction was not significant, $F(1, 367) = 2.5$, $p = .115$, $\eta_p^2 = 0.007$, a violation of the Transience Hypothesis. In fact, explicit attitudes trended in the opposite direction of the Transience Hypothesis: The effect of Plan Valence was non-significant in the load condition ($M_{\text{PositivePlan}}$ = 2.49, 95% CI [2.27, 2.71]; $M_{\text{NegativePlan}}$ = 2.38, 95% CI [2.17, 2.6]); $F(1, 367) = 0.46$, $p = .497$, $\eta_p^2 = 0.001$), and significant in the unload condition ($M_{\text{PositivePlan}}$ = 2.39, 95% CI [2.17, 2.61]; $M_{\text{NegativePlan}}$ = 1.94, 95% CI [1.72, 2.16]); $F(1, 367) = 8.31$, $p = .004$, $\eta_p^2 = 0.022$). Thus, as in Study 1 (and its replication; see Supplemental materials), explicit attitudes, but not implicit attitudes, violated the Transience Hypothesis.

8.4. Summary

We once again found support for our prediction that prepared reflexes shape attitudes, this time using a novel methodology. In a scenario distinct from the one in Studies 1–3, valenced action plans changed attitudes in a manner consistent with the Inaction, Dissociation, Transience, and Additivity hypotheses.

9. Study 5

The purpose of Study 5 was to demonstrate that prepared reflexes shape attitudes in more "real world" scenarios. We accomplished this by inducing prepared reflexes that associated non-human animals with a negative action representation (i.e. to hunt) or a positive action representation (i.e. to nurture). Whereas people rarely partake in criminal trials or executions, millions of people slaughter and/or nurture non-human animals on a daily basis.

9.1. Participants

Using MTurk, we recruited 257 native English speakers living in the United States. We excluded participants according to the same criteria as in the previous studies (N = 60), resulting in a final sample size of 197 (69% Female), ages 18 to 68 (M = 38, 95% CI [36, 40]). The exclusion rate (23%) did not differ by condition.

9.2. Procedure

We began the study by telling participants that we are interested in testing their ability to ignore images related to their current goals. To help us explore this question, our cover story explained, they would play a game called Homeland. Participants immediately proceeded to the first of two AMPs, which measured implicit attitudes toward two kinds of ducks, one with blue feathers and another with black feathers. We were interested in participants' implicit attitudes toward the blue ducks, and included black ducks as task-irrelevant controls. After collecting participants' baseline implicit attitudes toward blue ducks, we told participants that they would receive a $2 bonus if they win Homeland. We told all participants that their goal in Homeland is to provide food for their family. Participants then learned the rules of the game, with which we independently manipulated Plan Valence and Stimulus Valence, thus allowing us to test the Additivity Hypothesis.

In one of four scenarios, called the nurture scenario, we told participants that the most nutritious food in Homeland is the blue duck, which is quickly becoming endangered. Thus, to achieve their goal of feeding their family, participants in the nurture scenario had to grow the local population of blue ducks by nurturing them and protecting them from predators. Accordingly, in the nurture scenario, participants had a positive action plan toward blue ducks, and blue ducks were positively valenced. This is in contrast to another scenario, called the harvest scenario, in which participants learned that the most nutritious food in Homeland is the blue duck, and that they must slaughter as many blue ducks as possible in order to feed their family. Thus, in the harvest scenario, participants had a negative action plan toward blue ducks, and blue ducks were positively valenced.

In a third scenario, called the exterminate scenario, participants were told that the blue duck is an invasive species that is eliminating their primary food source. Thus, to achieve their goal of feeding their family, participants in the exterminate scenario had to slaughter as many blue ducks as possible. Accordingly, in the exterminate scenario, participants had a negative action plan toward blue ducks, and blue ducks were negatively valenced. Finally, we included a placate scenario to complete the factorial design. In the placate scenario, we told participants that the blue duck is an invasive species that is eliminating their primary food source, so, to achieve their goal of feeding their family, they must relocate the blue ducks to a place where they will be happier, healthier, and safer from predators. By helping the blue ducks, we explained, the blue ducks would stay away from participants' region and stop eliminating their food source. So, in the placate scenario, participants had a positive action plan toward blue ducks and blue ducks were negatively valenced.

To briefly summarize, these four scenarios reflect ecologically valid combinations of Plan Valence and Stimulus Valence. Throughout history, humans have nurtured (help + instrumental), harvested (harm + instrumental), exterminated (harm + non-instrumental) and placated (help + instrumental). Simulating these scenarios allows us to see whether prepared reflexes shape attitudes in ecologically valid contexts.

After participants learned the rules of Homeland, we asked them to indicate (1) whether they planned to help or harm the blue ducks and (2) whether the blue ducks were helpful or harmful to their goal of feeding their family. All participants answered both items correctly, indicating that we successfully manipulated both Plan Valence and Stimulus Valence. Next, participants completed the second AMP

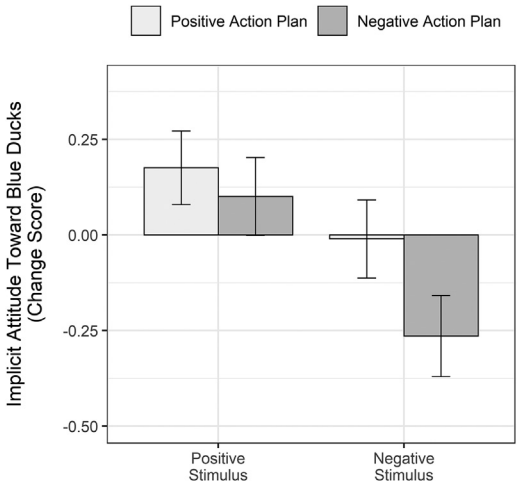

Fig. 5. Change in implicit attitudes toward blue ducks (after goal induction - before goal induction) as a function of Plan Valence (positive vs. negative) and Stimulus Content (unambiguously positive vs. unambiguously negative). Error bars represent 95% CIs.

measuring their implicit attitudes toward blue ducks relative to the task-irrelevant control ducks. As in all previous studies, participants never executed their action plan, thus allowing us to test the Inaction Hypothesis. To minimize the time and complexity of the procedure, we did not include an explicit measure of liking, nor did we include an Unload manipulation. Thus, the design of Study 5 allows us to test the Additivity Hypothesis and the Inaction Hypothesis, but not the Dissociation Hypothesis or the Transience Hypothesis.

9.3. Results

We ran a mixed ANOVA to assess the effects of Plan Valence, Stimulus Valence, and Time on implicit attitudes toward blue ducks (relative to control ducks). Consistent with the Additivity Hypothesis, we found a significant Plan Valence x Time interaction, $F(1, 193) = 10.09$, $p = .002$, $\eta_p^2 = 0.05$, and Stimulus Valence x Time interaction, $F(1, 193) = 28.4$, $p < .001$, $\eta_p^2 = 0.128$ (Fig. 5), but not a Stimulus Valence x Plan Valence x Time interaction, $F(1, 193) = 3.01$, $p = .084$, $\eta_p^2 = 0.015$. No other effects were significant.

Decomposing the Plan Valence x Time interaction, we found an effect of Plan Valence after plan formation, $F(1, 193) = 5.66$, $p = .018$, $\eta_p^2 = 0.028$, but not before. This effect was driven by increased implicit positivity after forming a positive action plan ($M = 0.07$, 95% CI [0.01, 0.14]) versus before ($M = -0.01$, 95% CI [−0.06, 0.04]), $F(1, 193) = 5.4$, $p = .021$, $\eta_p^2 = 0.027$, as well as by increased negativity after forming a negative action plan ($M = -0.04$, 95% CI [−0.11, 0.03]) versus before ($M = 0.04$, 95% CI [−0.02, 0.09]), $F(1, 193) = 4.72$, $p = .031$, $\eta_p^2 = 0.024$.

Decomposing the Stimulus Valence x Time interaction, we found an effect of Stimulus Valence after plan formation, $F(1, 193) = 30.47$, $p < .001$, $\eta_p^2 = 0.136$, but not before. This effect was driven by increased implicit positivity toward positively valenced blue ducks after plan formation ($M = 0.15$, 95% CI [0.08, 0.21]) versus before ($M = 0.01$, 95% CI [−0.04, 0.06]), $F(1, 193) = 15.07$, $p < .001$, $\eta_p^2 = 0.072$, as well as by increased implicit negativity toward negatively valenced blue ducks after plan formation ($M = -0.12$ 95% CI [−0.19, −0.05]) versus before ($M = 0.02$, 95% CI [−0.04, 0.07]), $F(1, 193) = 13.41$, $p < .001$, $\eta_p^2 = 0.065$.

9.4. Summary

The results of Study 5 are consistent with the Additivity and Transience hypotheses: valenced action plans that were never performed altered implicit evaluations irrespective of Stimulus Content. These results provide further support for the hypothesis that prepared reflexes shape attitudes, this time in the ecologically valid scenarios of nurturing, harvesting, exterminating, and placating.

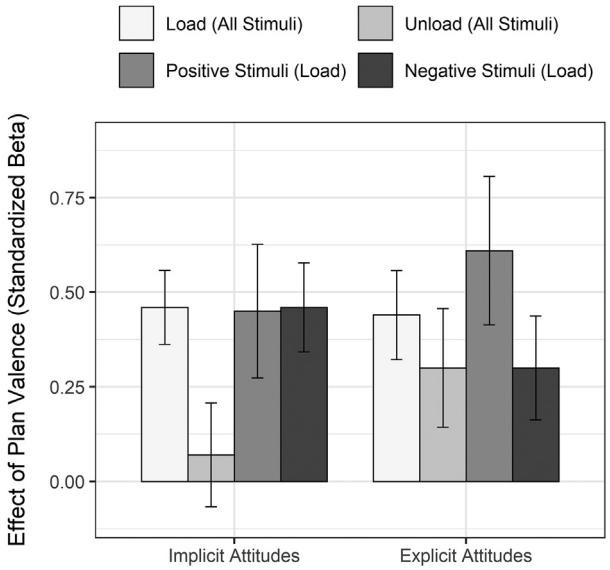

Fig. 6. Standardized betas denoting effect sizes of Plan Valence on attitudes as a function of Attitude Type (implicit vs. explicit) and Unload (load vs. unload). Effect sizes within load conditions are further broken down by Stimulus Content (unambiguously positive vs. unambiguously negative). Error bars represent 95% CIs.

10. Meta-analysis

To estimate the cumulative support for our proposal that prepared reflexes shape attitudes, we conducted a meta-analysis. We submitted data from all 6 studies to a linear mixed effects regression predicting evaluations of the target stimulus. This regression included random intercepts for subject and study, as well as fixed effects for Stimulus Content, Unload, Plan Valence, Attitude Type, and their interaction terms (see Fig. 6).

The fixed effect of Attitude Type was a binary factor denoting whether the attitude was implicit or explicit. To permit direct comparisons between implicit and explicit attitudes, we standardized scores on the two respective measures, and modeled implicit attitudes at Time 2 only.

This model supported all four key hypotheses. Consistent with the Inaction Hypothesis, we found that, in the load conditions, Plan Valence had a significant effect on implicit attitudes ($b = 0.46$, SE = 0.05 $t(2814.87) = 8.63$, $p < .001$) despite the fact that participants never executed their action plans. Consistent with the Transience Hypothesis, we found that, for implicit attitudes, the Unload x Plan Valence interaction was significant ($b = 0.39$, $SE = 0.09$, $t(2680.17) = 4.27$, $p < .001$): Plan Valence shaped implicit attitudes in the load conditions (as we have seen), but not in the unload conditions ($b = 0.07$, $SE = 0.08$, $t(2701.22) = 0.92$, $p = .36$). Consistent with the Additivity Hypothesis, we found that, in the load conditions, the Stimulus Content x Plan Valence interaction for implicit attitudes was not significant ($b = 0.004$, $SE = 0.11$, $t(2832.57) = 0.04$, $p = .968$): Plan Valence had equivalent, significant effects on implicit attitudes toward unambiguously positive ($b = 0.45$, $SE = 0.09$, $t(2832.55) = 5.21$, $p < .001$) and negative stimuli ($b = 0.46$, $SE = 0.06$, $t(2832.8) = 7.25$, $p < .001$).

For the Dissociation Hypothesis, we found two sources of support. First, we found a significant Attitude Type x Unload x Plan Valence interaction ($b = 0.25$, $SE = 0.11$, $t(1385.78) = 2.23$, $p = .026$). As we

have seen, the Unload x Plan Valence interaction was significant for implicit attitudes, but for explicit attitudes, it was non-significant (b = 0.15, SE = 0.10, t(2738.14) = 1.53, p = .127): Plan Valence had a significant effect on explicit attitudes in both the load (b = 0.44, SE = 0.06, t(2912.46) = 7.23, p < .001) and unload conditions (b = 0.30, SE = 0.08, t(2701.02) = 3.91, p < .001). Thus, unlike implicit attitudes, explicit attitudes violated the Transience Hypothesis.

We also found that, in the load conditions, the Attitude Type x Stimulus Content x Plan Valence interaction was significant (b = 0.32, SE = 0.15, t(1540.22) = 2.11, p = .035). As we have seen, the Stimulus Content x Plan Valence interaction was non-significant for implicit attitudes, but for explicit attitudes, it was significant (b = 0.31, SE = 0.12, t(2920.41) = 2.52, p = .012): Plan Valence had a weaker effect on explicit attitudes toward unambiguously negative (b = 0.30, SE = 0.07, t(2888.92) = 4.38, p < .001) versus positive stimuli (b = 0.61, SE = 0.10, t(2930.37) = 5.9, p < .001). Thus, unlike implicit attitudes, explicit attitudes violated the Additivity Hypothesis.

11. General discussion

Why do people like what they like and dislike what they dislike? The results of 6 studies suggest a novel answer: *prepared reflexes*. When people plan to perform a positively valenced action R+, their (implicit) attitude toward the target stimulus S becomes more positive due to an S-R+ association in working memory; likewise, when people plan to perform a negatively valenced action R-, their (implicit) attitude toward S becomes more negative due to an S-R- association in working memory. In line with this view, we found that valenced action plans shaped (implicit) attitudes in a manner consistent with four hypotheses whose joint validity is uniquely consistent with the operation of prepared reflexes: The Inaction Hypothesis, the Transience Hypothesis, the Additivity Hypothesis, and the Dissociation Hypothesis (see Table 1).

Consistent with the Inaction Hypothesis, valenced action plans changed attitudes even though they were never performed. Consistent with the Transience Hypothesis, valenced action plans changed (implicit) attitudes only for as long as the prepared reflexes that corresponded to those action plans were in working memory. Consistent with the Additivity Hypothesis, we found that (implicit) attitudes changed despite the fact that the target person's valence was unambiguous, irrespective of whether the target person's valence was positive or negative, and even if the target person was Adolf Hitler. Consistent with the Dissociation Hypothesis, we found diverging patterns of implicit and explicit attitude change such that the former, but not the latter, was consistent with the operation of prepared reflexes: Unlike implicit attitudes, explicit attitudes violated both the Transience Hypothesis and the Additivity Hypothesis. As we have seen (Table 1), this constellation of effects is consistent with the operation of prepared reflexes, but not with alternative mechanisms by which goals change attitudes, including cognitive dissonance, self-perception, AA-training, and biased scanning.

11.1. On the relationship between implicit attitudes and behavior

The present findings have important implications for one of the most pressing questions in the field of implicit social cognition: Under what conditions do implicit attitudes predict behavior? The importance of this question is illustrated by the results of a recent meta-analysis conducted by Kurdi et al. (2019, see also Amodio, 2018; Payne, Vuletich, & Lundberg, 2017). These researchers estimated the correlation between implicit attitudes and behavior (i.e. implicit-criterion correlations, or ICCs), as well as the relationship between ICCs and "conceptual moderators" (i.e. variables that should, according to prominent theories of implicit cognition, predict ICCs). Although implicit attitudes were reliable predictors of behavior, none of the conceptual moderators predicted ICCs. Reflecting on this surprising result, the authors noted that "The absence of theoretical predictors of ICCs… suggests that theorizing about implicit cognition is relatively unsophisticated at this time." (p. 14).

Why is it so hard to predict whether and to what extent implicit attitudes will predict behavior? One reason, we suspect, is that an implicit attitude created by a prepared reflex will predict only that behavior which the prepared reflex elicits; such an implicit attitude reflects a particular S-R association, and thus should only predict R. One cannot reliably predict the behavioral outcomes of such an implicit attitude without knowing the content of the prepared reflex from which it emerged.

Consider the following example. An experimenter explores whether implicit attitudes predict behavior toward homeless people. Her strategy is to bring participants to the lab, measure their implicit attitudes toward homeless people, and then assess behavior. Although the experimenter does not know it, half her participants plan to volunteer at a soup kitchen that weekend. These participants have prepared reflexes that associate homeless people with the positive act of serving warm meals, and thus have relatively positive implicit attitudes toward homeless people during the experimental session. The remaining participants in her study do not plan to volunteer at the soup kitchen, not because they like homeless people any less than the others, but because they had other plans that weekend. These participants do not have prepared reflexes associating homeless people with a positive action, and thus their implicit attitudes toward homeless people are relatively negative. In this scenario, the experimenter should find a correlation between implicit attitudes and behavior if and only if she operationalizes behavior as volunteering at a soup kitchen that weekend; planning to volunteer at a soup kitchen that weekend is, in this scenario, the only source of variance in implicit attitudes toward homeless people. If the experimenter were to measure some other behavior – financial donations to homeless shelters, for instance – no effect would emerge.

As this example illustrates, when prepared reflexes create implicit attitudes, those implicit attitudes may predict only a highly restricted range of specific and idiosyncratic behaviors. This may shed light on the critically important questions of why ICCs are so difficult to predict, and how this issue may be resolved. Specifically, it suggests that one must determine the contents of people's prepared reflexes in order to make reliable predictions about the behavioral outcomes of their implicit attitudes.

11.2. On the interface between attitudes and action control

More broadly, the present findings illustrate how the action control literature can advance our understanding of attitudes. Given the strong link between systems underlying action and evaluation, this literature can generate novel hypotheses about attitude formation and change, building on those explored here. One such hypothesis bears on the self-regulation of implicit bias: People can dramatically reduce their implicit bias simply by relabeling their intended actions toward outgroups.

This idea stems from work exploring how action representations become positively or negatively valenced. As we have said, one such process is brute force association: the more an action is paired with positive or negative valence, the more positive or negative that action's representation becomes (e.g., Eder, Rothermund, De Houwer, & Hommel, 2015). But action valence also depends on how actions are labeled. By relabeling one's intended action, one can change that action's valence while holding its other features (e.g., motor commands and outcomes) constant. For instance, the act of pushing a computer joystick can be labeled *push away* or *push upwards* (Eder & Rothermund, 2008). Though the action is identical either way, its representation is more positive when labeled *upwards* versus *away*, because upwards is a more positive concept (Eder & Rothermund, 2008). This finding, combined with the present results, suggests that by relabeling their intended actions, people can radically alter their implicit biases from one

moment to the next.

To illustrate, consider a White person preparing to interact with a Black person. The White person may form either of the following action plans: “Avoid acting prejudiced” or “Make my interaction partner feel valued and respected”. Though these plans may entail the same set of behaviors, the latter has a more positive label, and thus should have a more positive representation. By linking this relatively positive action to the Black interaction partner, the plan to “Make my interaction partner feel valued and respected” should, through the operation of a prepared reflex, reduce the White person's implicit bias.

Consistent with this, Trawalter and Richeson (2006) found that participants expended less cognitive effort during an interracial interaction if their action plan had a positive label (i.e. promote a positive interracial interaction) versus a negative label (i.e. avoid prejudice). This finding was interpreted in terms of the self-regulatory benefits of adopting a so-called promotion focus (Higgins, 1997, 1998), but theories of action control suggest another (not mutually exclusive) interpretation: through the operation of prepared reflexes, the positively labeled action plan induced more positive attitudes toward the interaction partner, which made it easier to interact successfully. On this view, the positive label did not just help people control the expression of their racial bias — it actually reduced their racial bias by transferring the positivity of an action representation to a representation of an outgroup member. More broadly, this view suggests that how people label their intended actions may shape the outcomes of their social interactions by dramatically altering their evaluations of others. Such an account is consistent with the emerging consensus that evaluation involves not just the retrieval of attitudes from memory, but also the construction of attitudes on the fly (Albarracin & Shavitt, 2018; Payne et al., 2017). At the same time, it illustrates the generativity of action control literature for the study of attitudes.

11.3. Concluding remarks

Ever since Thurstone (1928) declared that “attitudes can be measured,” psychologists have sought to illuminate the factors that shape people's likes and dislikes. The present findings further this aim by revealing a new mechanism of attitude change that bears on both the nature and predictive power of human likes and dislikes. In doing so, they invite experts in the domains of attitudes and action control to form a closer alliance, one that mirrors the bond between the phenomena they study.

Open practices

Studies 1–6 in this article earned Open Materials and Open Data badges for transparent practices. Materials and data for all studies are available at https://osf.io/ytkcf.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103950.

References

Albarracin, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology, 69*, 299–327.

Amodio, D. M. (2018). Social Cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences, 23*(1), 21–33.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38*(9), 1194–1208.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(1), 1–23.

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? *Advances in experimental social psychology. 20. Advances in experimental social psychology* (pp. 65–122). Elsevier.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323.

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology. 6. Advances in experimental social psychology* (pp. 1–62). Elsevier.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current Directions in Psychological Science, 26*(5), 422–428.

Brendl, C. M., & Higgins, E. T. (1996). Principles of judging valence: What makes events positive or negative? *Advances in experimental social psychology. 28. Advances in experimental social psychology* (pp. 95–160). Elsevier.

Cabanac, M. (1971). Physiological role of pleasure. *Science, 173*(4002), 1103–1107.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond Bipolar Conceptualizations and Measures: The Case of Attitudes and Evaluative Space. *Personality and Social Psychology Review, 1*(1), 3–25.

Centerbar, D. B., & Clore, G. L. (2006). Do approach-avoidance actions create attitudes? *Psychological Science, 17*(1), 22–29.

Cohen, A.-L., Bayer, U. C., Jaudas, A., & Gollwitzer, P. M. (2008). Self-regulatory strategy and executive control: Implementation intentions modulate task switching and Simon task performance. *Psychological Research, 72*(1), 12.

Cohen-Kdoshay, O., & Meiran, N. (2009). The representation of instructions operates like a prepared reflex: Flanker compatibility effects found in first trial following S–R instructions. *Experimental Psychology, 56*(2), 128–133.

Cole, M. W., Braver, T. S., & Meiran, N. (2017). The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. *Neuroscience & Biobehavioral Reviews, 81*, 4–15.

Collins, B. E., & Hoyt, M. F. (1972). Personal responsibility-for-consequences: An integration and extension of the “forced compliance” literature. *Journal of Experimental Social Psychology, 8*(6), 558–593.

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

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8*(7), 342–353.

Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best? *Developmental Cognitive Neuroscience, 4*, 38–51.

Eder, A. B., & Hommel, B. (2013). Anticipatory control of approach and avoidance: An ideomotor approach. *Emotion Review, 5*(3), 275–279.

Eder, A. B., & Klauer, K. C. (2007). Common valence coding in action and evaluation: Affective blindness towards response-compatible stimuli. *Cognition and Emotion, 21*(6), 1297–1322.

Eder, A. B., & Klauer, K. C. (2009). A common-coding account of the bidirectional evaluation–behavior link. *Journal of Experimental Psychology: General, 138*(2), 218.

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis) match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General, 137*(2), 262.

Eder, A. B., Rothermund, K., De Houwer, J., & Hommel, B. (2015). Directive and incentive functions of affective action consequences: An ideomotor approach. *Psychological Research, 79*(4), 630–649.

Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology, 67*(3), 382.

Exner, S. (1879). Physiologie der Grosshirnrinde. In L. Hermann (Vol. Ed.), *Handbuch der physiologie. vol. 2. Handbuch der physiologie* (pp. 189–350). Leipzig, Germany: Vogel.

Fagot, C. (1994). *Chronometric investigations of task switching* (Ph.D.)San Diego: University of California.

Fazio, R. H. (1987). *Self-perception theory: A current perspective.* (Paper presented at the Social influence: the Ontario symposium).

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition, 25*(5), 603–637.

Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology, 87*(5), 557.

Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row Peterson.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203.

Fishbach, A., Shah, J. Y., & Kruglanski, A. W. (2004). Emotional transfer in goal systems. *Journal of Experimental Social Psychology, 40*(6), 723–738.

Fitzsimons, G. M., & Fishbach, A. (2010). Shifting closeness: Interpersonal effects of personal goal progress. *Journal of Personality and Social Psychology, 98*(4), 535.

Fitzsimons, G. M., & Shah, J. Y. (2008). How goal instrumentality shapes relationship evaluations. *Journal of Personality and Social Psychology, 95*(2), 319.

Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology, 1*(2), 145–155.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review, 15*(4), 352–366.

Fujita, K., & Carnevale, J. J. (2012). Transcending temptation through abstraction: The role of construal level in self-control. *Current Directions in Psychological Science, 21*(4), 248–252.

Fujita, K., Trope, Y., Liberman, N., & Maya, L.-S. (2006). Construal Levels and Self-Control. *Journal of Personality and Social Psychology, 90*(3), 351–367.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in experimental social psychology. 44. Advances in experimental social psychology* (pp. 59–127). Elsevier.

Gawronski, B., & Brannon, S. M. (2018). What is cognitive consistency and why does it matter? In E. Harmon-Jones (Ed.). *Cognitive dissonance: Progress on a pivotal theory in social psychology*(2 ed.). Washington, DC: American Psychological Association.
Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40*(4), 535–542.
Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493.
Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*(1), 1.
Haidt, J. (2003). The moral emotions. *Handbook of affective sciences. 11(2003). Handbook of affective sciences* (pp. 852–870).
Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*(12), 1280.
Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. *Advances in experimental social psychology. 30. Advances in experimental social psychology* (pp. 1–46). Elsevier.
Hommel, B. (2000). 11 The prepared reflex: Automaticity and control in stimulus-response translation. *Control of cognitive processes* (pp. 247). .
Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negativity Information Weighs More Heavily on the Brain: The Negativity Bias in Evaluative Categorizations. *Journal of Personality and Social Psychology, 75*(4), 887–900.
Jones, C. R., Vilensky, M. R., Vasey, M. W., & Fazio, R. H. (2013). Approach behavior can mitigate predominately univalent negative attitudes: Evidence regarding insects and spiders. *Emotion, 13*(5), 989.
Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology, 92*(6), 957.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480.
Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284.
Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ... Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569.
Lavender, T., & Hommel, B. (2007). Affect and action: Towards an event-coding account. *Cognition and Emotion, 21*(6), 1270–1296.
Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., & Schacter, D. L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science, 12*(2), 135–140.
McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model. *Dual-process theories of the social mind* (pp. 204–217). .
Meiran, N. (2000). Modeling cognitive control in task-switching. *Psychological Research, 63*(3–4), 234–249.
Meiran, N. (2005). Task rule-congruency and Simon-like effects in switching between spatial tasks. *The Quarterly Journal of Experimental Psychology Section A, 58*(6), 1023–1041.
Meiran, N., Cole, M. W., & Braver, T. S. (2012). When planning results in loss of control: Intention-based reflexivity and working-memory. *Frontiers in Human Neuroscience, 6*, 104.
Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences, 201714945*.
Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406–19415.
Miles, J. D., & Proctor, R. W. (2008). Improving performance through implementation intentions: Are preexisting response biases replaced? *Psychonomic Bulletin & Review, 15*(6), 1105–1110.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
Neumann, R., Förster, J., & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and evaluation. *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 371–391). .
Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*(5), 653.
Olson, J. M., & Stone, J. (2005). The influence of behavior. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.). *Handbook of attitudes and attitude change* (pp. 223–271). Mahwah, NJ: Erlbaum.
Orehek, E., & Forest, A. (2016). When People Serve as Means to Goals: Implications of a Motivational Account of Close Relationships. *Current Directions in Psychological Science, 25*(2), 79–84.
Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin, 39*(3), 375–386.
Payne, B. K., Cheng, S. 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.
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.
Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248.
Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380.
Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in experimental social psychology. vol. 20. Advances in experimental social psychology* (pp. 297–340). Elsevier.
Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences, 114*(32), 8511–8516.
Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.
Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995.
Sénémeaud, C., & Somat, A. (2009). Dissonance arousal and persistence in attitude change. *Swiss Journal of Psychology, 68*(1), 25–31.
Simon, H. (1982). *Models of bounded rationality*. Cambridge, MA: MIT Press.
Sinclair, L., & Kunda, Z. (1999). Reactions to a black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*(5), 885.
Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social Tuning of Automatic Racial Attitudes: The Role of Affiliative Motivation. *Journal of Personality and Social Psychology, 89*(4), 583–592.
Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131–142.
Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. *Advances in experimental social psychology. 11. Advances in experimental social psychology* (pp. 249–288). Elsevier.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529–554.
Trawalter, S., & Richeson, J. A. (2006). Regulatory focus and executive function after interracial interactions. *Journal of Experimental Social Psychology, 42*(3), 406–412.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Van Dessel, P., Eder, A. B., & Hughes, S. (2018). Mechanisms underlying effects of approach-avoidance training on stimulus evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(8), 1224.
Van Dessel, P., Hughes, S., & De Houwer, J. (2018). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review, 1088868318795730*.
Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science, 22*(4), 490–497.
Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*(5), 815.
Woodworth, R. S. (1938). *Experimental psychology*. New York, NY: Holt, Rinehart and Winston.
Woud, M. L., Maas, J., Becker, E. S., & Rinck, M. (2013). Make the manikin move: Symbolic approach–avoidance responses affect implicit and explicit face evaluations. *Journal of Cognitive Psychology, 25*(6), 738–744.