22 Implicit Person Memory: Domain-General and Domain-Specific Processes of Learning and Change[1]

Benedek Kurdi[1] *and Mahzarin R. Banaji*[2]
[1]*Yale University*
[2]*Harvard University*

Imagine that you learn that a co-worker of yours recently rear-ended a car. How do you update your impression of this person? Does it make a difference whether you observe the accident directly, or you hear about it from a friend? Or perhaps from the office gossip who has been known to spread rumors even if they lack any basis in reality? Does the gender or the race of the co-worker make any difference? What if this is the third time that the same co-worker has caused an accident? And what if you also know that the road was icy each time?

Person memory,[2] a subdiscipline of social cognition research, is synonymous with the birth of social cognition as a field of study. Work on person memory emerged from a small conference in the late 1970s, organized by a group of social psychologists who recognized a new opportunity to advance their field by using existing methods of cognition, especially measures of explicit memory, to study the structure and organization of knowledge about other humans (Hastie et al., 1980).

Person memory researchers investigated whether and how humans would update their beliefs about other humans when confronted with new knowledge about them, including in relatively mundane cases (such as driving ability), and also considerably more complex ones. In each study, a situation involving a single individual would be presented, with experimental designs that included multiple conditions varying information about the person and the context. Judgments of the target individual and memory for the presented information served as the main dependent measures. At its core, person memory as a field took up questions of how knowledge about other individuals is acquired, stored, retrieved, and updated (e.g., Higgins et al., 1977; Kerpelman & Himmelfarb, 1971; Srull & Wyer, 1979; Winter & Uleman, 1984).

Along with person memory's focus on the individual, another subfield of social cognition, social group cognition, also made rapid progress. Here the focus was on representations of individuals as members of social groups, such as age, gender, race/ethnicity, and sexuality. Interestingly, person memory and social group cognition remained, to a large degree, theoretically and

DOI: 10.4324/9781003045687-25

methodologically independent of each other, in spite of their close conceptual connection. Notably, from the earliest days, the idea of automaticity took center stage in the study of social group cognition; as such, measures of implicit rather than explicit memory were adapted to investigate group-based attitudes and stereotypes.

Among the first such paradigms used by social cognition researchers was sequential priming, a procedure originally designed to explore the organization of conceptual knowledge in human memory (Meyer & Schvaneveldt, 1971; Neely, 1976). These methods were then specifically adapted to the study of social categories, such as gender and race, with a focus on both implicit attitudes and stereotypes (Devine, 1989; Dovidio et al., 1986; Fazio et al., 1986; Gaertner & McLaughlin, 1983). The goal of demonstrating the presence of automatic attitudes and stereotypes, both generally and in intergroup contexts more specifically, characterized early research (Banaji et al., 1993; Banaji & Hardin, 1996). Presumably due to the dominant theoretical view that implicit social group cognition was resistant to new information (Bargh, 1999; Devine, 1989; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Wilson et al., 2000), most relevant research tended to stay away from questions of change, or the acquisition and updating of representations given new knowledge. Notable exceptions emphasized the goal-dependent nature of implicit social cognitive processes, including implicit stereotyping (Moskowitz, 1996; Moskowitz et al., 1999).

In this chapter, we focus on a body of work that has used implicit measures, such as sequential priming (Fazio et al., 1986, 1995), the Implicit Association Test (Greenwald et al., 1998), the Affect Misattribution Procedure (Payne et al., 2005), and their variants, to study how evaluations of and beliefs about individual human targets are acquired and how they shift in the face of new information. As such, these studies provide insights into phenomena and processes of *implicit person memory*, i.e., knowledge about individuals that is retrieved under conditions of automaticity. By use of the term, we do not mean to suggest that any of the authors whose work we discuss below would have subscribed to this label themselves or, importantly, that their work would have been guided by a shared set of theoretical assumptions. In fact, one of the conclusions emerging from this brief overview points to the need for a stronger focus on theory building if research on implicit person memory is to make progress.

It is odd, in retrospect, that these two lines of research, one on person memory and the other on social group cognition, both pursued similar goals of applying measures of memory to the study of social entities, and yet operated in parallel, with little cross-fertilization. The fact that one engaged with explicit forms of memory and the other with implicit forms of memory hardly seems to be a sufficient reason for these lines of research to have remained separate, with little to no cross-talk. This is not to say that important exceptions did not exist. Notably, work by Uleman and colleagues as well as Moskowitz and colleagues on spontaneous trait inferences was devoted to the study of automatic processes in person memory from the earliest days of the field (e.g., Moskowitz, 1993; Moskowitz & Roman, 1992; Uleman et al., 1996; Winter & Uleman, 1984; see also Newman, 1991). Similarly, Bargh and colleagues used subliminal priming to investigate whether the accessibility of certain constructs (including traits such as honesty or hostility) could influence processes of learning and judgment about individual social targets even outside of conscious awareness (e.g., Bargh & Pietromonaco, 1982; Bargh & Thein, 1985; Pratto & Bargh, 1991).

Research on implicit social group cognition and person memory connected in more profound and far-reaching ways in the mid-2000s when papers on implicit person memory started to appear in larger numbers (Castelli et al., 2004; DeCoster et al., 2006; McConnell et al., 2008; Meersmans et al., 2005; Rydell et al., 2006). From person memory, the study of implicit person memory inherited its core question—an interest in knowledge about individual humans; from implicit social group cognition, it inherited its core method—an emphasis on memory and judgment that occurs in automatic form.

The remainder of this chapter is structured as follows. In the first part of the chapter, we review what is already known about implicit person memory. For the sake of clarity and as a first tentative step toward theory building, we present existing implicit person memory research as belonging to one of two basic categories.

In the first category, we discuss implicit person memory work that does not emphasize the uniquely human nature of human targets or the importance of uniquely social processes of reasoning. Instead, such work uses human targets incidentally to explore how implicit evaluations and beliefs are acquired and how they change. In doing so, this research does not assume that processes of acquisition or change differ depending on the targets of learning. Rather, the tacit understanding underlying these experiments seems to be that products and brands, the self, social categories, abstract concepts, significant others, or political parties are fundamentally interchangeable with each other and with single individuals as the targets of learning. This subset of implicit person memory work emphasizes questions about the inputs to and the processes contributing to attitude and belief acquisition and change. For example, among the inputs investigated are approach/avoidance training, evaluative conditioning, and verbal statements of different kinds. When it comes to process, much attention has been devoted to the distinction between association formation mechanisms registering merely that two stimuli go together in the environment and propositional processes also encoding the specific types of relationships that stimuli can share with each other.

In the second category, we review studies that have investigated implicit person memory by attempting to identify processes specific to learning about social targets. The themes emerging from this subset of implicit person memory work include the interplay between individual-level and category-level information in implicit attitude acquisition and change, the role of facial cues, diagnostic narrative information, and the reinterpretation of previously encountered behavioral evidence about a person. This latter body of experimental work, which operates under an assumption of the uniqueness of social learning processes, raises the complementary theoretical issue of whether these inputs and mechanisms are, in fact, unique to the social domain.

At a first glance, the approaches taken by these two sets of studies seem intrinsically incompatible: Learning about novel social targets cannot at the same time be essentially equivalent to learning about a brand or an abstract idea and also fundamentally different from it. Competing assumptions of domain-general vs. domain-specific processes in human learning and memory are, of course, not specific to the study of implicit person memory; rather, they are ubiquitous across social psychology and the cognitive sciences more broadly. Domain-general accounts posit that the computations characterizing human cognition are fundamentally the same no matter whether someone is thinking about Reese Witherspoon, the number line, or high-calorie foods (e.g., Banaji & Bhaskar, 2000; Ruff & Fehr, 2014); meanwhile, domain-specific theories suggest that human thought cannot be properly characterized without adequately considering the type of object that the person is thinking about (e.g., Cosmides & Tooby, 1994; Sperber, 1994).

Against this general theoretical backdrop as well as the apparent contradiction between the two sets of empirical studies reviewed above, we devote considerable space in the third (and final) section of this chapter to the issue of whether a domain-specific account of implicit person memory is worth proposing and defending. We also address other important topics that are yet to be settled in this area. These topics include differing definitions of what it means for a learning process to be effective, conditions of encoding, and probably the thorniest issue of all: the content and format of the mental representations mediating implicit person memory and, more generally, implicit social cognition.

Implicit Person Memory as a Case Study of Domain-General Processes

Early implicit social group cognition work inherited from studies of conceptual organization in the human mind the method of sequential priming (e.g., Meyer & Schvaneveldt, 1971; Neely, 1976). In sequential priming studies, researchers measure participants' speed and accuracy in responding to a target (e.g., the word "butter") after exposure to different primes, some assumed to be semantically related to the target (e.g., the word "bread") and some assumed to be unrelated (e.g., the word "democracy").

Along with the sequential priming paradigm, early associative theories of implicit social cognition (e.g., Devine, 1989; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Wilson et al., 2000) also adopted the theoretical framework commonly used to interpret findings from this paradigm: spreading activation models of semantic memory (Collins & Loftus, 1975). These models assume that the human mind encodes concepts (such as "butter", "bread", "good", "calculating", "African American", and "democracy") via a set of nodes in a vast semantic network. The closer two concepts are associated with each other in meaning, the stronger the connections between them in the network and, as such, the more likely encountering one is to automatically co-activate the representation of the other. The strength of connections, in turn, is assumed to be driven by something akin to a simple form of associative learning, or the Hebbian principle of activity-dependent synaptic plasticity whereby concurrent firing of neurons strengthens their connection—the idea captured by the mnemonic "what fires together wires together" (Hebb, 1949). It then follows that concepts frequently encountered in close temporal and spatial proximity (such as "butter" and "bread") will come to be strongly connected, whereas concepts infrequently or never encountered together (such as "butter" and "democracy") will be relatively weakly, or not at all, connected.

Importantly, in the early days of implicit social cognition research, the dual assumptions of (a) associative representations and (b) low-level, trial-by-trial associative learning seemingly obviated the need to study the acquisition and change of implicit evaluations and beliefs in the lab. After all, if implicit attitudes merely reflect the piecemeal shift of associative strengths in response to the long-term co-occurrence statistics of the environment, then lab-based learning paradigms may not be particularly informative for at least two reasons. First, learning processes were assumed to be too mechanical and simple to be worth studying at all. Second, relatively minor manipulations of the kind implemented in the lab were not expected to be impactful in shifting a lifetime of experience tracking co-occurrences.

However, by the mid-2000s, theoretical work in implicit social cognition emancipated itself from spreading activation models of memory and, importantly, from the assumption of purely associative learning giving rise to implicit evaluations and beliefs. Notably, the associative–propositional evaluation (APE) model by Gawronski and Bodenhausen (2006) still assumed that implicit evaluations are subserved by conceptual associations stored in long-term memory. At the same time, it also began to stake out the idea that these associations can be sensitive not only to co-occurrences experienced in the environment but also, indirectly, to the relational content of propositions.[3] That is, they are assumed to encode not only the fact that two stimuli are associated with each other and the degree of their relatedness but also the type of relationship that they share with each other. For instance, under strict associative theories, exposure to statements such as "Donald is not delusional" is expected to produce an ironic effect of strengthening the connection between the conceptual nodes "Donald" and "delusional" in long-term memory. By contrast, under the APE model, at least in certain cases, implicit evaluations can reflect the propositional content of the statement, thus strengthening the conceptual connection between "Donald" and "rational" rather than the purely associative "Donald" and "delusional".

Later, De Houwer and his colleagues formulated an even more radical proposal (e.g., De Houwer, 2007, 2014; Mitchell et al., 2009), which has since gained much empirical traction. Specifically, they posited that associative processes of learning and representation are not necessary to account for the acquisition and change of implicit evaluations at all. Rather, similar to their explicit counterparts, implicit evaluations were assumed to be able to shift quickly and dynamically (rather than only in response to vast numbers of stimulus co-occurrences). This idea represented a radical departure from previous thinking according to which implicit cognition taps associative structures and is therefore immune to propositional reasoning. As such, propositional theories went further in expanding the scope of potential inputs to implicit social cognition than even the most flexible dual-process accounts available at the time, such as the APE model mentioned above. Moreover, propositional accounts did away with the idea of associative representation. Instead, implicit and explicit evaluations were both thought to emerge from propositional representations (e.g., "Donald is delusional") and assumed to differ only in terms of the conditions of their retrieval. Specifically, propositional accounts suggest that implicit evaluations are characterized by relatively more automatic and explicit evaluations by relatively more controlled processes of activating the same type of propositional knowledge stored in long-term memory.

These new theoretical developments have fueled innovative empirical work on the acquisition and change of implicit evaluations and beliefs for at least three reasons. First, the APE model and, to a considerably larger extent, propositional accounts, popularized the idea that implicit social cognition may be amenable to the same basic processes of flexible updating as explicit social cognition, including its propensity for quick and dynamic revision in the face of relational information. If this is the case, then those processes of updating needed to be explored experimentally. Second, with theoretical disagreement between associative accounts, dual-process accounts, and propositional accounts regarding the processes of learning and representation underlying implicit evaluation came the desire to advance the debate and to reach a satisfactory resolution. Third, the APE model and propositional accounts are both yet to be formulated with sufficient computational specificity to derive falsifiable predictions from them. As such, efforts to constrain these theories using empirical data, and to eventually develop versions specific enough to be falsifiable, have been ongoing ever since these accounts were first introduced. Given interest in and methods available for formal modeling of mental processes, we are cautiously optimistic about the likelihood of success at this time.

Against this theoretical backdrop, a considerable number of implicit person memory studies have attempted to answer two distinct but related questions. First, what types of input are capable of producing change in implicit evaluations of or beliefs about novel human targets? Researchers have studied different types of input that can roughly be divided into the following categories: (a) approach/avoidance training (e.g., Van Dessel et al., 2015, 2016); (b) attribute conditioning, that is, repeated pairings of a target with stimuli related to a semantic category (e.g., Förderer & Unkelbach, 2015, 2016); (c) evaluative conditioning, that is, repeated pairings of a target with intrinsically valenced stimuli (e.g., Förderer & Unkelbach, 2013;

Gast & Rothermund, 2011a, 2011b; Rydell & Jones, 2009); and (d) behavioral statements (e.g., Boucher & Rydell, 2012; Cone et al., 2019, 2021; Moran et al., 2015, 2017; Olcaysoy Okten et al., 2019; Olcaysoy Okten & Moskowitz, 2020; Peters & Gawronski, 2011; Rydell et al., 2007). Second, what kind of learning processes mediate learning via the different manipulations mentioned previously? Specifically, are learning processes uniquely sensitive to associative information (co-occurrences of stimuli in the environment), or do they also encode relational information (the different types of relationship that those stimuli can share with each other)?

Responsiveness of Implicit Person Memory to Different Types of Learning

The first major finding emerging from this literature, which seems robust if not incontrovertible given the strength of the evidence, is that implicit person memory is flexible (that, is capable of changing) in the face of a variety of different inputs, including the types of information described above. Such inputs include approach/avoidance training, attribute conditioning, evaluative conditioning, and behavioral statements. For example, participants in the studies by Van Dessel et al. (2015) updated implicit evaluations of novel human targets that they approached in a positive direction and those that they avoided in a negative direction. Likewise, participants in the studies by Förderer and Unkelbach (2015) updated implicit beliefs of targets on trait dimensions such as athleticism as a result of repeated pairings of the targets with material semantically related to those trait dimensions. Participants have also been shown to adjust implicit evaluations of targets paired with positive stimuli in a positive direction and those paired with negative stimuli in a negative direction (Rydell & Jones, 2009). Finally, Rydell et al. (2007) found that participants revised their implicit evaluations of targets in both positive and negative directions in a lawful manner in response to verbal statements. Taken together, this body of work demonstrates that implicit evaluations of and beliefs about novel human targets are subject to change, including in response to relatively minimal experimental manipulations.

This basic result, which has been replicated dozens of times, seems fundamentally incompatible with the idea that implicit evaluations and beliefs require vast amounts of information to form and then to change. After all, the experiments referenced above involved exposure to information about novel individuals for relatively short periods of time ranging from a few minutes to no more than an hour. Arguably, this time frame is insufficient for the types of protracted learning processes posited by traditional associative theories (e.g., Devine, 1989; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Wilson et al., 2000) to be essential for implicit attitude acquisition and change to unfold. By contrast, this body of evidence is considerably easier to reconcile with more flexible dual-process theories (e.g., Gawronski & Bodenhausen, 2006) and single-process propositional theories (e.g., De Houwer, 2007, 2014;

Mitchell et al., 2009), which allow for the possibility that implicit evaluations could be updated dynamically in the face of relatively small amounts of information.

The Role of Associative vs. Propositional Processes in Implicit Person Memory

At the same time, it is notable that at least three of the four types of manipulations described above, including approach/avoidance training, attribute conditioning, and evaluative conditioning, are commonly assumed to be associative in at least two senses of the word. First, these paradigms create learning via repeated co-occurrences of a target with a stimulus or action. Second, they are usually thought to reflect the products of such learning by strengthening conceptual associations in long-term memory. Arguably, the behavioral statements used in the paradigms described above (such as "Mike cheated during a poker game") can also be interpreted in associative terms given that they include co-occurrences of a target (e.g., "Mike") with valenced words (e.g., "cheat"; Caliskan et al., 2017; Kurdi & Dunham, 2021). As such, based on these results alone, it seems that the only minor change required to make traditional associative theories compatible with the data on learning and change is to allow for the possibility that associative learning can unfold quickly, perhaps after as few as a dozen trials or even in response to a single, highly potent, stimulus pairing. This possibility is by no means incompatible with theories and empirical findings on associative learning from outside the social cognition context (e.g., Drew et al., 2010; Gershman, 2015; Rescorla & Wagner, 1972).

However, another finding, now also broadly replicated, appears to be even more fundamentally incompatible with a purely associative notion of implicit person memory (for reviews, see Cone et al., 2017; De Houwer et al., 2020; Kurdi & Dunham, 2020). Specifically, under associative accounts, implicit evaluations and beliefs are thought to reflect exclusively the fact that two things go together in the environment and the number of times that they have been paired with each other. However, in direct contradiction to this idea, implicit evaluations and beliefs have been robustly demonstrated to also reflect *how* two pieces of information are related to each other.

Here we mention only a few cases in which implicit evaluations were found to encode relational information in a way that seems fundamentally incompatible with associative accounts. For example, implicit evaluations of novel targets in the studies by Peters and Gawronski (2011) and Boucher and Rydell (2012) were sensitive to whether the content of statements about those targets was affirmed or negated: A person presented along with the behavior "continually yells at his wife in public" was evaluated negatively when the behavior was revealed to be characteristic of him; however, when it was revealed to be uncharacteristic, implicit evaluations shifted in a positive direction. The idea that abstract knowledge of this kind would be crucial, or even relevant, to implicit person memory would have been difficult to entertain in a predominantly associative framework. Notably, Kurdi and Dunham (2021) even found that the updating of implicit evaluations of a novel target depended on whether participants made normative errors in propositional inference, such as denying the antecedent, providing further evidence for the importance of high-level reasoning processes. Together, these results strongly suggest that associative processes alone are insufficient to account for the patterns of learning and updating observed in implicit person memory.

Domain-Specific Processes in Implicit Person Memory

The implicit person memory studies reviewed in the previous section share the important commonality that they have been designed to test relatively domain-general theories of implicit social cognition. These theories assume, more or less tacitly, that processes of learning and representation cut across different types of human (and even non-human) targets and that, therefore, different types of target stimuli used to investigate such processes are relatively interchangeable with each other. In fact, in our own work, we have conducted learning studies involving existing social categories and non-social targets (e.g., Kurdi & Banaji, 2017), novel social groups (e.g., Kurdi & Dunham, 2021), and novel individuals (e.g., Mann et al., 2020) without systematically investigating whether these targets differ from each other in theoretically relevant ways. Nevertheless, convergent results obtained across different categories of stimuli suggest that the underlying learning process is sufficiently general to produce similar outcomes. Such a result may be seen as surprising from the perspective of theories across the cognitive sciences that have emphasized the importance of domain-specific processes to human learning and memory (e.g., Cosmides & Tooby, 1994; J. P. Mitchell et al., 2005; Saxe & Kanwisher, 2003; Sperber, 1994).

By contrast, the studies discussed in this section have focused on inputs to and processes of implicit attitude acquisition and change that are relatively specific to the domain of person memory. Such domain specificity is usually related to one of two aspects of studies: the types of information being presented and the types of information processing assumed to occur. Of course, these two aspects are intertwined with each other more often than not, but here we discuss each of them separately for ease of presentation.

On the one hand, some studies have relied on information about novel social targets that would not be meaningfully interpretable outside the social domain. Such studies have included experiments probing the interplay of individual-level and category-level information (e.g., Cao & Banaji, 2016; Gawronski et al., 2003; McConnell et al., 2008; Rubinstein et al., 2018; Rubinstein & Jussim, 2019) and the effects of facial cues on implicit evaluation (e.g., Gawronski & Quinn, 2013; Shen et al., 2020). On the other hand, studies have also presented information to participants that was assumed to give rise to domain-specific processes of social reasoning: diagnostic information about a target's true moral character (e.g., Cone et al., 2019, 2021; Cone & Ferguson, 2015) or information prompting participants to reinterpret a target's previously encountered behaviors (e.g., Kurdi et al., 2021b; Mann & Ferguson, 2015, 2017; Olcaysoy Okten et al., 2019).

Additional Evidence for Flexibility and the Role of Relational Information

The studies reviewed here differ from the studies reviewed previously in their emphasis on uniquely social types of information and inference. However, at the same time, similar to the relatively domain-general studies discussed previously, they can provide evidence on the flexibility of implicit person memory in the face of different types of input as well as on the role of relational information in processes of updating. Indeed, these relatively domain-specific studies have provided ample evidence for the flexible updating of implicit evaluations. As such, their findings largely converge with the experiments relying on relatively domain-general information reviewed above in suggesting that, contrary to influential early conceptualizations of implicit social cognition as resistant to updating (e.g., Bargh, 1999; Devine, 1989; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Wilson et al., 2000; but see Moskowitz et al., 1999; Blair, 2002), implicit person memory is remarkably flexible in response to ever-changing informational inputs.

Moreover, similar to the set of domain-general studies reviewed above, updating in many of these experiments seems to have unfolded in a way that is difficult to reconcile with notions of slow and piecemeal associative learning to the exclusion of propositional processes of reasoning, which is a central assumption of associative accounts of implicit social cognition. For example, Cone and Ferguson (2015) demonstrated that implicit evaluations of a novel target formed from dozens of behavioral statements can reverse in valence from positive to negative as a result of a single piece of highly diagnostic information about that target (e.g., the person having mutilated a small, defenseless animal). Given the extremity of the valence of the novel information, this result in and of itself may be explained by a particularly potent form of associative learning. However, the finding that the strength of updating tracked the extent to which participants believed that the novel information was believable and diagnostic of the target's moral character (Cone et al., 2019, 2021) seems even more fundamentally incompatible with purely associative accounts.

Studies relying on the idea of reinterpretation have produced findings that are similarly difficult to reconcile with piecemeal association formation mechanisms. In these studies, unlike in most work on the updating of implicit evaluations, attitude change is not achieved by presenting entirely novel information about the target; rather, participants are prompted to reconsider the evaluative implications of already known information. For example, Kurdi et al. (2021b) have shown that exposure to excerpts from a real-world podcast, containing a mix of positive and negative information, can lead to considerable updating of initially highly negative evaluations of an individual. Furthermore, similar to the studies relying on diagnostic information reviewed above, the amount of updating was predicted by the extent to which participants found the novel counterattitudinal information persuasive. Such ubiquitous involvement of higher-order reasoning processes in implicit person memory seems difficult if not impossible to reconcile with the idea of a cognitive system that merely tracks co-occurrences of targets with valenced information in the environment.

Evidence of Domain-Specific Processes

Notably, given their reliance on certain types of (social) information and (social) reasoning, these experiments also additionally inform about relatively domain-specific processes of implicit person memory. These insights concern the relative importance of category-level and individual-level information, use of facial cues, and the role of diagnostic information and reinterpretation. As mentioned in the introduction, we see an inherent contradiction between the approaches of (a) treating human targets as fundamentally interchangeable with other classes of stimuli (such as brands or abstract concepts) in implicit person memory work vs. (b) assuming that implicit evaluations of and beliefs about human targets are sensitive to a unique set of inputs and learning processes. As such, we hope that placing these two groups of studies side by side and critically reviewing both sets of underlying assumptions will prove helpful in reaching a resolution and achieving theoretical integration.

Category-Level Information vs. Individuating Information

A relatively large number of studies have investigated the interplay between and relative importance of category-level information (e.g., information that a target is a man or Iranian American) and individuating information (e.g., information that they rear-ended a car or took money from a donation box) in implicit person memory. Given the uniquely social nature of both the social category information and the individuating information used in these studies, this work can reasonably be interpreted as informing about domain-specific processes of social learning and memory.

In an early study, Gawronski et al. (2003) demonstrated that implicit evaluations of social categories can bias the process of forming an impression of individuals belonging to those categories. Specifically, participants in these studies interpreted ambiguous behaviors performed by a Black target more negatively than the same ambiguous behaviors performed by a White target but only to the extent that they had relatively positive implicit attitudes toward White Americans and relatively negative implicit attitudes toward Black Americans. Given the uniquely social nature of both the category-level and individual-level information used by Gawronski et al. (2003), this experiment seems to provide early evidence for the involvement of uniquely domain-specific processes in implicit person memory.

More recent work has investigated the formation of implicit evaluations of and implicit beliefs about novel social targets more directly by presenting category-level and individual-level information to participants that were contradictory in their evaluative or semantic implications. Similar to the Gawronski et al. (2003) study, given the uniquely social nature of both types of information, these experiments are broadly assumed to inform about domain-specific inputs to implicit person memory.

For example, Cao and Banaji (2016) introduced participants to a male and a female target (category-level information) and then described the former as a nurse and the latter as a doctor (individuating information). Although implicit beliefs shifted significantly relative to baseline, they were still indicative of the persistence of stereotype-congruent associations of the female target with the category "nurse" and the male target with the category "doctor." Other work using different designs and different targets has produced results ranging from complete reliance of implicit person memory on category-level information to the exclusion of individual-level information (McConnell et al., 2008) to complete reliance on individual-level information to the exclusion of category-level information (Rubinstein et al., 2018; Rubinstein & Jussim, 2019). As such, further empirical and conceptual work will be necessary to reconcile these seemingly contradictory findings with each other. However, crucially, as a set, these studies seem to provide compelling evidence for the role of uniquely social types of input in the updating of implicit evaluations and beliefs.

The Role of Facial Information

A second, considerably smaller, set of studies have investigated the influence of the human face on implicit person memory. The effects of different facial cues, such as the shape of the face, the distance between the eyes, and the height of the forehead, on impression formation have been well documented using self-report measures (for reviews, see Todorov et al., 2008, 2015). Crucially from our perspective, similar to social category information, facial features are widely seen as a source of uniquely social information. As such, studies investigating the effects of facial cues on implicit measures of evaluation and belief can also provide information on relatively domain-specific mechanisms of learning and change in implicit person memory.

In a first relevant study by Gawronski and Quinn (2013), participants read positive and negative behavioral statements about novel targets (presented as faces) and then completed implicit measures of attitude toward previously unseen targets whose faces were manipulated to appear similar to the targets about whom participants had learned earlier. Implicit evaluations generalized to these novel target faces, thus providing initial evidence for the idea that facial cues can influence implicit person memory. In more recent work, Shen et al. (2020) produced a conceptually similar finding, demonstrating that targets whose faces were manipulated to appear extremely untrustworthy engendered highly negative implicit evaluations. At the same time, diagnostic behavioral information about the same targets (see below) led to the revision, and sometimes full reversal, of the face-based negative evaluations. Again, these studies seem to provide evidence for the operation of relatively domain-specific processes of learning and updating in implicit person memory.

Diagnostic Information and Reinterpretation

Research by Ferguson, Cone, Mann, and colleagues has investigated in detail two seemingly uniquely social forms of updating in implicit person memory: the first relying on diagnostic information (Cone et al., 2019, 2021; Cone & Ferguson, 2015) and the second on the reinterpretation of previously encountered behavioral information (Kurdi et al., 2021b; Mann et al., 2020; Mann & Ferguson, 2015, 2017).

As alluded to earlier, the first type of paradigm tends to pit two types of information against each other: a large number of behavioral statements implying a positive evaluation of a novel target and a single piece of extremely negative and diagnostic behavioral information about the same target. Notably, these studies are theoretically well integrated with, and have directly expanded upon, a long line of work relying on explicit measures of impression formation (e.g., Reeder & Brewer, 1979; Reeder & Coovert, 1986; Trafimow & Schneider, 1994). Specifically, they show that negative behavioral information, especially extremely negative behavioral information, tends to give rise to particularly strong dispositional inferences and that these inferences, in turn, influence not only explicit but also implicit evaluations. Given that dispositional inferences are widely seen as uninterpretable outside a social context, these studies can also be construed as providing evidence for the operation of uniquely social processes in implicit person memory.

Similar to studies involving diagnostic behavioral information, studies relying on the idea of reinterpretation are also usually assumed to demonstrate the flexibility of implicit evaluations in the face of uniquely social information. The typical design of reinterpretation experiments involves presenting an initial narrative that is rich in negative episodic details (e.g., Mann & Ferguson, 2015, 2017; but see Olcaysoy Okten et al., 2019). For example, participants in Mann and Ferguson (2015) were introduced to a novel target called Francis West and read a relatively long vignette about him breaking into his neighbors' homes to remove "precious things" from them. Based on this initial information, participants construed West's actions as a burglary and evaluated him negatively on both explicit and implicit measures. Subsequently, in the reinterpretation condition, participants learned that West entered the houses because they were on fire and the "precious things" that he removed (saved) were actually the neighbors' children. Although the second piece of information is minimal in length and detail compared with the narrative presented at the outset of the study, it was sufficient to induce revisions to, and often full reversals of, the initially formed negative evaluations. As such, this line of work provides additional evidence for the effective and rapid revision of implicit evaluations in the face of relatively domain-specific forms of reasoning about social information.

Interim Summary: The Flexibility of Implicit Person Memory

To summarize the insights gained from the work reviewed previously, evidence for the possibility of rapidly and dynamically revising implicit evaluations of novel social targets seems overwhelming. Processes of revision can unfold in response to information that could be characterized as relatively domain-general (including actions to approach or avoid targets, pairings of targets with intrinsically pleasant or unpleasant stimuli, and exposure to valenced verbal descriptions of targets), or in response to information that could be characterized as more specific to social targets (including competing category-level and individuating information, facial cues, diagnostic behavioral information, and information giving rise to the reinterpretation of previously encountered behaviors). These inputs to the updating of implicit evaluations clearly go beyond simple stimulus pairings; moreover, learning can emerge highly effectively, within a matter of a few minutes. Finally, even learning from seemingly simple paradigms involving the repeated presentation of stimulus pairings has been shown to be modulated by the meaning with which participants imbue those stimulus pairings, either spontaneously or as a result of relational information provided by the experimenter.

Overall, these results are difficult to reconcile both with most early conceptualizations of implicit social cognition, inherited from spreading activation models of memory, as well as the associative accounts building on these early conceptualizations (e.g., Bargh, 1999; Devine, 1989; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Wilson et al., 2000; but see Moskowitz et al., 1999; Blair, 2002). After all, according to these accounts, implicit evaluations and beliefs can be updated only as a result of protracted learning involving vast numbers of stimulus pairings. Moreover, under these theories, implicit cognition is thought to be sensitive exclusively to co-occurrence information experienced in the environment without reflecting the ways in which such information is construed by the reasoner. By contrast, these findings are compatible with propositional accounts of implicit evaluation (e.g., De Houwer, 2007, 2014; Mitchell et al., 2009) as well as other theories that do not posit a strict separation between an implicit system reflecting an associative mode of processing and an explicit system reflecting a propositional mode of processing (e.g., Cunningham et al., 2007; Fazio, 2007; Gawronski & Bodenhausen, 2006; Kurdi & Dunham, 2020).

In summary, the body of knowledge generated by the field of implicit person memory (and implicit social cognition more broadly) over the past two decades has led to a fundamental revision of most early conceptualizations of implicit attitudes. Specifically, there has been a significant movement away from theories uniquely emphasizing the importance of associative processes of learning and representation toward theories emphasizing (additionally or exclusively) the importance of propositional processes of learning and representation. Although substantial disagreement still remains about the relative contributions of different types of processes to implicit evaluation (e.g., Gawronski & Bodenhausen, 2018; De Houwer, 2018; Kurdi & Dunham, 2020; McConnell & Rydell, 2014), we believe that this shift alone demonstrates the considerable promise of the experimental approaches taken since the mid-2000s in unraveling the nature of implicit cognition and social learning.

Open Empirical and Theoretical Questions in Implicit Person Memory

In spite of the theoretically rich insights that have emerged over the past 20 years of research on implicit person memory, a sizable number of issues, some of them of crucial theoretical importance, are yet to be resolved or even to be systematically addressed. Some of these issues are specific to implicit person memory but several of them apply, *mutatis mutandis*, to implicit social cognition more generally. In the remainder of the chapter, we offer a brief and subjective overview of these unresolved issues. We hope that this overview will provide an impetus for new theory development and empirical work, or at least serve as a basis for discussions about what directions new theoretical and empirical approaches should take.

Do We Need a Domain-Specific Theory of Implicit Person Memory?

The apparent inconsistency in basic theoretical assumptions between the two sets of studies reviewed above seems in need of resolution. Specifically, although these theoretical assumptions are rarely if ever discussed explicitly, the first set of studies, relying on paradigms such as approach/avoidance training, attribute conditioning, evaluative conditioning, and verbal statements, seem to assume that the processes and mechanisms of implicit evaluation are largely domain-general. From this assumption it follows that (a) paradigms originally developed in the context of animal learning, and only later adapted to the study of human cognition and human social cognition, are generally well-suited to the study of implicit attitude acquisition and change and (b) the stimuli used in these paradigms (be they non-words, shapes, products, social groups, abstract concepts, or single individuals) are generally interchangeable with each other. Virtually all theories of implicit evaluation discussed in this chapter so far seem to make exactly the same assumption simply by virtue of being silent on the possibility of any domain-specific processes of implicit evaluation.

By contrast, the second set of studies seem to make a fundamentally different assumption, namely that there are at least some processes of implicit person memory that cannot be described using domain-general mechanisms. These studies have provided evidence for learning unfolding via the interplay of individuating and social category information, facial cues, diagnostic behavioral information, and reinterpretation. Notably, to the degree that these studies are embedded in existing theoretical frameworks, they tend to emphasize accounts that have been formulated in the context of person memory, such as the continuum model of impression formation (Fiske & Neuberg, 1990), interpersonal transference (Chen & Andersen, 1990), and attribution theories (e.g., Reeder & Brewer, 1979).

As presently implemented, these two approaches seem fundamentally incompatible with each other. Implicit evaluations of individual targets cannot be subserved by exclusively, or at least mostly, domain-general processes and also, simultaneously, by exclusively, or at least mostly, domain-specific processes. At least three possibilities for resolving this apparent inconsistency are worth considering. First, it is conceivable that the processes underlying implicit evaluation are mostly domain-general. Second, it is conceivable that the processes underlying implicit evaluation are mostly domain-specific. Finally, it is also conceivable that this level of analysis is too coarse and different aspects of implicit evaluation should be investigated separately along the continuum from fully domain-general to fully domain-specific.

Without prejudging how these open questions will be resolved, we believe that substantial amounts of theoretical work and empirical evidence already exist to suggest that these issues are sufficiently important to be experimentally addressed. First, the most critical takeaway from propositional theories of implicit evaluation (e.g., De Houwer, 2007, 2014; Mitchell et al., 2009) and empirical work informed by these theories is that associative processes relying exclusively on the idea of registering co-occurrence information are insufficient to capture the processes of acquisition and change observed in experimental studies. However, if this is the case, and processes of high-level reasoning have a ubiquitous influence on implicit evaluation, then the possibility that humans might reason differently about social and non-social entities, and even different social entities, cannot be dismissed out of hand. The amount of existing research that has shown distinctions between social and non-social processing, especially when measures of neural activation are included (e.g., J. P. Mitchell et al., 2005; Saxe & Kanwisher, 2003; Young et al., 2007), strongly suggests that this issue should become a priority for testing.

As an example from recent theoretical work, Faure et al. (2020) have made a compelling case for studying implicit evaluation in the context of close relationships and for incorporating the findings from such studies into overarching theories of implicit social cognition. Specifically, these authors point out that these theories tend to treat race attitudes (and, to some degree, other intergroup attitudes) and attitudes toward novel experimental targets as ideal typical cases of implicit evaluation. However, unlike these two attitude domains, highly consequential implicit evaluations of close others, such as family members and romantic partners, stem from rich, complex, and constant, or at least repeated, personal experience that fluctuates in both valence and intensity. If the processes giving rise to implicit evaluations of close others and other social entities differ from each other in such a clear and potentially consequential way, then beliefs and evaluations in other domains that are being routinely investigated using implicit measures of (social) cognition, such as consumer goods and brands (Dimofte, 2010), addictive substances (Lindgren et al., 2020), and other attitude objects relevant to psychopathology (Teachman et al., 2019), may also differ considerably from each of those areas and also from each other.

Moreover, a systematic comparative approach would presumably also prompt investigators to be more precise about the types of differences that may exist between social and non-social attitude objects that are relevant to processes of attitude acquisition and revision. For example, the studies investigating the effects of conflicting social group information and individuating information seem to tacitly assume that this distinction is uniquely relevant to human targets. Although certain aspects of such information may indeed be unique, cases of contradictory category-level and individual-level information can also be considered a specific instance of the effects of how categorical and exemplar-specific information are integrated with each other in long-term memory. This issue has been studied extensively across different subfields of psychology (e.g., Medin et al., 1984; Merriman et al., 1997; Schapiro et al., 2017).

Similarly, although studies investigating the updating of implicit evaluations via diagnostic behavioral information and reinterpretation seem to assume that reasoning on the basis of these two types of input is uniquely social, such reasoning may at least in part be supported by domain-general processes. For example, the diagnostic behavioral information used in studies by Cone, Ferguson, and colleagues tends to be both negative and extreme in valence and, as such, the more general phenomenon of negativity dominance (Rozin & Royzman, 2001) may contribute to the effect. Of course, one could make the argument that the mechanisms underlying updating of implicit evaluations via diagnostic behavioral information cannot be satisfactorily explained by simple negativity dominance because the effect does not arise if the target is only incidentally associated with negative information (Cone & Ferguson, 2015). However, the process of assigning observable outcomes in the world to hidden latent causes is by no means specific to the social domain (Gershman et al., 2015; Gershman & Niv, 2010). Moreover, Kurdi et al. (2021c) have provided tentative evidence for the operation of negativity dominance specifically in the context of the acquistion of non-social attitudes.

A potentially defining difference between implicit person memory and other forms of implicit evaluation may be the sensitivity of the former, but not of the latter, to reasoning about hidden mental states, such as goals, beliefs, and desires (J. P. Mitchell et al., 2005; Saxe & Kanwisher, 2003;

Young et al., 2007). For example, Saxe and Kanwisher (2003) have found differences in neural activation in response to scenarios involving false beliefs (e.g., Sally erroneously believing that an object is in one box rather than in the other) vs. scenarios involving other types of false representations (e.g., an outdated photograph, such as a photograph of an apple hanging from a tree, which has been blown to the ground by a strong wind since the picture was taken). Similarly, Mitchell et al. (2005) identified unique patterns of neural activation when participants were asked to consider targets' psychological states (e.g., "curious" or "energetic") rather than their physical features.

In line with the idea that mental state reasoning may be an important contributor to implicit person memory, a recent line of studies by Kurdi et al. (2020) have provided evidence for the sensitivity of implicit evaluations to targets' accurate and false beliefs about the world. Specifically, in these studies, implicit evaluations of novel targets were more negative when they caused positive rather than negative outcomes. However, holding the valence of the outcome constant, implicit evaluations deviated more strongly from neutrality when the outcomes were caused intentionally (e.g., putting poison in someone's coffee knowing that it is poison) rather than unintentionally (e.g., putting poison in someone's coffee erroneously believing that it is sugar).

However, although these results are suggestive, there remains much to be explored about the potential uniqueness of some types of input to, or processes underlying, implicit person memory. First, the vignettes used in the experiments by Kurdi et al. (2020) featured extremely negative outcomes. As such, the results may not generalize to more mundane cases, which would undercut the idea that implicit person memory universally involves mental state reasoning. Second, whether mental state reasoning consistently contributes to implicit evaluations of social targets is unclear. It is possible that mental state reasoning may be impactful only in cases where it is directly applicable to the problem at hand (such as diagnostic behavioral information). Alternatively, such reasoning may operate by default even in cases where seemingly no relevant information is being provided (such as evaluative conditioning). Third, not all human targets activate mental state reasoning to the same degree (e.g., Harris & Fiske, 2006; McLoughlin & Over, 2017), which may make the ubiquity of this potentially uniquely social input to the updating of implicit evaluations questionable.

What about Different Definitions of Effectiveness?

The overwhelming majority of studies reviewed in this chapter equate the effectiveness of different forms of person learning, at least tacitly, with their immediate capacity to modulate responding on an implicit measure of belief or evaluation. However, as suggested by recent investigations in the context of implicit race attitudes, the temporary malleability of implicit evaluations need not translate into enduring change (Lai et al., 2014, 2016). A handful of studies have already explored, and provided evidence for, the durability of change in implicit person memory brought about by different interventions (Cone et al., 2021; Kurdi et al., 2021a; Mann et al., 2020; Mann & Ferguson, 2017; Ranganath & Nosek, 2008). However, we hope that such investigations will become more commonly implemented in the future given that they have the potential to inform both about basic mechanisms of learning and change and about the effectiveness of different interventions in producing long-term shifts in implicit evaluation.

Moreover, if learning is to be truly effective, it should generalize across different contexts. In the study of both animal learning (e.g., Bouton & Bolles, 1979; Bouton & Peck, 1992) and human memory (e.g., Schiller et al., 2010, 2013) it is commonplace to assume that novel information contradicting the evaluative implications of a prior learning episode does not usually erase the memory trace associated with the original experience. Rather, it tends to create a new memory trace, which now competes with the old memory trace for expression. Although a relatively large body of evidence exists to suggest that implicit evaluations can be context-dependent (Gawronski et al., 2018), the boundary conditions and replicability of such effects are not sufficiently well understood (Gawronski et al., 2015). Moreover, with the sole exception of a study by Brannon and Gawronski (2017), none of these studies have systematically investigated the relative context (in-)dependence of different types of input to implicit evaluation.

Finally, next to nothing is known about whether and to what degree implicit evaluations created via different types of knowledge differ from each other in terms of their resistance to novel counterattitudinal information (but see Kurdi et al., 2021a). In fact, related to the conclusions of the previous section, different criteria of relative effectiveness may not yield consistent results across different areas of attitude acquisition and change. As such, based on presently available data, it can be confidently concluded that processes of implicit person memory (and implicit social cognition more broadly) are momentarily malleable in the face of complex information that goes well beyond simple co-occurrence information, including reasoning about causes and effects, mental states, and the believability and diagnosticity of the evidence that one encounters. A small number of studies additionally suggest that, at least in the context of novel social targets, such effects can persist beyond a single experimental session. However, very little is known about the (relative) context-specificity of different learning modalities as well as the resistance of their outputs to countervalent information.

What about Different Encoding Conditions?

Similar to the dearth of research on different conditions under which newly learned information about social targets can be retrieved, the state of knowledge about the effects of different encoding conditions on processes of updating in implicit person memory is extremely limited. In most experimental studies reviewed above, and in most studies on the acquisition and change of implicit evaluations more generally, participants are able and motivated to focus on the evaluative information presented to them. Moreover, they are usually specifically instructed to memorize the information to which they are exposed. A few studies have manipulated the availability of cognitive resources during learning (e.g., Boucher & Rydell, 2012; Mann & Ferguson, 2015; Shen et al., 2020); however, these studies have yielded conflicting results. Notably, recent work by Fan et al. (2021) suggests that when cognitive resources are available, implicit evaluations can spontaneously reflect the effects of relational information; however, under cognitive load, implicit attitudes seem to be uniquely sensitive to co-occurrence information. These results call into question the ubiquity of propositional influences on implicit evaluation and highlight the need for further inquiry into the boundary conditions of such effects.

Moreover, the problem may run too deep to be solved simply by placing participants under cognitive load while they are exposed to information designed to create or shift implicit evaluations. Remarkably, based on recent studies by Wimmer and Poldrack, the results of single-session learning studies in which participants encounter novel information in a highly massed fashion may not at all, or only under extremely limited conditions, generalize to more ecologically valid settings in which reasoners are usually exposed to information about the same target across multiple occasions over time (Wimmer et al., 2018; Wimmer & Poldrack, 2021).

Specifically, these experiments suggest that when information is presented in a massed way to participants, the effectiveness of even model-free processes of value-based learning, long assumed to be emerging in a purely stimulus-driven way, is highly correlated with individual differences in working memory capacity. However, when the same information was administered to participants across three occasions over a three-day period, the correlation between working memory capacity and value-based learning disappeared entirely. As such, based on these results, it seems premature to conclude that truly associative processes cannot contribute to evaluative learning (Corneille & Stahl, 2018). Rather, the paradigms routinely used to try to produce such effects may simply not create the appropriate psychological conditions for those very effects to emerge.

Finally, What about Mental Representations?

Based on the evidence reviewed previously, it is quite clear that implicit evaluations can (at least momentarily) reflect the effects of different types of relational information in a way that accounts relying on association formation mechanisms alone cannot explain. However, propositional accounts of implicit evaluation (e.g., De Houwer, 2007, 2014; Mitchell et al., 2009) are committed to a considerably stronger theoretical claim—namely, that implicit attitudes emerge from the automatic activation of propositional representations. At present, there is no evidence to suggest that this more sweeping idea is accurate, either in implicit person memory or beyond.

Specifically, as discussed in more detail in Kurdi and Dunham (2020), explicit and implicit evaluations could be thought of as being sensitive to the same basic sources of information, including relational information, but encoding such information at different levels of compression. To take an example from the person memory domain, let's assume that individuals can differ from each other along three dimensions: warmth, competence, and physical attractiveness. Each individual receives a score on each of the three dimensions (akin to a probabilistic implementation of truth values) and a weighted sum of the three scores is used to calculate the overall evaluation.

In this setting, explicit evaluations may be conceptualized as encoding the scores on each dimension, the weights, as well as the resulting summary evaluation, whereas implicit evaluations may be conceptualized as encoding only the resulting summary evaluation without having access to the specific pieces of (propositional) information from which the summary evaluation emerged. In fact, some initial evidence obtained mainly in non-social contexts seems to suggest that the idea of compression, entirely absent from both dual-process and propositional theories, can be useful in understanding some patterns of flexibility and recalcitrance in the updating of implicit evaluations (e.g., Kurdi et al., 2019, 2021c). This foray notwithstanding, the issue remains open for further exploration.

Conclusion

In this chapter, we provided a brief overview of a field we refer to as implicit person memory. Implicit person memory encompasses experimental studies investigating the acquisition (learning) and revision (updating) of implicit attitudes toward and implicit beliefs about novel social targets in response to different types of information. The evidence reviewed here seems to provide unequivocal support for the immediate malleability of implicit evaluations in the face of multiple sources of information, some of which are routinely regarded as emerging from domain-general mechanisms (e.g., evaluative conditioning) and others of which are routinely regarded as emerging from domain-specific processes of social reasoning (e.g., mental state inferences). These results are difficult to reconcile with most early conceptualizations of implicit evaluation as (a) purely associative and (b) generally resistant to updating.

The effects of relational information on implicit person memory are extremely well-established, and support for the possibility of rapid revisions to implicit attitudes is equally robust. However, considerably less is known about (a) the domain-specific vs. domain-general nature of the processes by which implicit evaluations are updated; (b) the generalizability of and mechanisms underlying updating across different domains; (c) the persistence of updating effects over time, their context-specificity, and the resistance of updating to counterattitudinal information; (d) the scope of encoding conditions under which implicit evaluations can exhibit sensitivity to relational information; and, finally, (e) the mental representations mediating the effects of co-occurrence and relational information on implicit evaluation.

We hope that, by summarizing available evidence on these issues and by highlighting gaps in the existing literature, the present review will help create a coherent and systematic theory of implicit person memory and a more comprehensive, accurate, and easily falsifiable theory of implicit social cognition.

Notes

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

[2] The term "person perception" is currently used considerably more frequently than the term "person memory." However, the term can cause confusion when used in psychology broadly because, with few exceptions, "person perception" research does not investigate any truly perceptual processes. Therefore, and to maintain connection with the history of the field, we have opted to use the term "person memory," although it may seem anachronistic to some readers.

[3] As such, the APE model is not a fully propositional account of implicit evaluation but rather a flexible dual-process model that, like propositional accounts, allows for a role of propositional processes but, unlike propositional accounts, retains the idea of associative learning and representation from early theories of implicit social cognition.

References

Banaji, M. R., & Bhaskar, R. (2000). Implicit stereotypes and memory: The bounded rationality of social beliefs. In D. L. Schacter, & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 139–175). Harvard University Press.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141. 10.1111/j.1467-9280.1996.tb00346.x

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*(2), 272–281. 10.1111/j.1467-9280.1996.tb00346.x

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). The Guilford Press.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, *43*(3), 437–449. 10.1037/0022-3514.43.3.437

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology*, *49*(5), 1129–1146. 10.1037/0022-3514.49.5.1129

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6(3), 242–261. 10.1207/s15327957pspr0603_8

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin*, *38*(10), 1329–1342. 10.1177/0146167212450464

Bouton, M. E., & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*(4), 445–466. 10.1016/0023-9690(79)90057-2

Bouton, M. E., & Peck, C. A. (1992). Spontaneous recovery in cross-motivational transfer (counter conditioning). *Animal Learning & Behavior*, *20*(4), 313–321. 10.3758/bf03197954

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context-(in)dependent updating of implicit evaluations. *Social Psychological and Personality Science*, *8*(3), 275–283. 10.1177/1948550616673875

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. 10.1126/science.aal4230

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, *113*(27), 7475–7480. 10.1073/pnas.1524268113

Castelli, L., Zogmaister, C., Smith, E. R., & Arcuri, L. (2004). On the automatic evaluation of social exemplars. *Journal of Personality and Social Psychology*, *86*(3), 373–387. 10.1037/0022-3514.86.3.373

Chen, S., & Andersen, S. M. (1990). Relationships from the past in the present: Significant-other representations and transference in interpersonal life. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 123–190). Academic Press. 10.1016/s0065-2601(08)60273-7

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. 10.1037/0033-295x.82.6.407

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

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

Cone, J., Flaharty, K., & Ferguson, M. J. (2021). The long-term effects of new evidence on implicit impressions of other people. *Psychological Science*. Advance online publication. 10.1177/0956797620963559

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In James M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 56, pp. 131–199). Elsevier. 10.1016/bs.aesp.2017.03.001

Corneille, O., & Stahl, C. (2018). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161–189. 10.1177/1088868318763261

Cosmides, L., & J. Tooby (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld, & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 85–116). Cambridge University Press.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*(5), 736–760. 10.1521/soco.2007.25.5.736

DeCoster, J., Banner, M. J., Smith, E. R., & Semin, G. R. (2006). On the inexplicability of the implicit: Differences in the information provided by implicit and explicit tests. *Social Cognition*, *24*(1), 5–21. 10.1521/soco.2006.24.1.5

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*(2), 230–241. 10.1017/s1138741600006491

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

De Houwer, J. (2018). A functional–cognitive perspective on the relation between conditioning and placebo research. *Neurobiology of the Placebo Effect Part I*, *138*, 95–111. Elsevier. 10.1016/bs.irn.2018.01.007

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 61, pp. 127–183). Elsevier. 10.1016/bs.aesp.2019.09.004

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

Dimofte, C. V. (2010). Implicit measures of consumer cognition: A review. *Psychology & Marketing*, *27*(10), 921–937. 10.1002/mar.20366

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22–37. 10.1016/0022-1031(86)90039-9

Drew, M. R., Denny, C. A., & Hen, R. (2010). Arrest of adult hippocampal neurogenesis in mice impairs single- but not multiple-trial contextual fear conditioning. *Behavioral Neuroscience*, *124*(4), 446–454. 10.1037/a0020081

Fan, X., Bodenhausen, G. V., & Lee, A. Y. (2021). Acquiring favorable attitudes based on aversive affective cues: Examining the spontaneity and efficiency of propositional evaluative conditioning. *Journal of Experimental Social Psychology*. Advance online publication. 10.1016/j.jesp.2021.104139

Faure, R., McNulty, J. K., Hicks, L. L., & Righetti, F. (2020). The case for studying implicit social cognition in close relationships. *Social Cognition*, *38*(Supplement), s98–s114. 10.1521/soco.2020.38.supp.s98

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

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. 10.1037//0022-3514.69.6.1013

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

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology* (Vol. 21, pp. 1–74). 10.1016/s0065-2601(08)60317-2

Förderer, S., & Unkelbach, C. (2013). On the stability of evaluative conditioning effects: The role of identity memory, valence memory, and evaluative consolidation. *Social Psychology*, *44*(6), 380–389. 10.1027/1864-9335/a000150

Förderer, S., & Unkelbach, C. (2015). Attribute conditioning: Changing attribute assessments through mere pairings. *The Quarterly Journal of Experimental Psychology B*, *68*(1), 144–164. 10.1080/17470218.2014.939667

Förderer, S., & Unkelbach, C. (2016). Changing US attributes after CS–US pairings changes CS attribute assessments. *Personality and Social Psychology Bulletin*, *42*(3), 350–365. 10.1177/0146167215626705

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*(1), 23–30. 10.2307/3033657

Gast, A., & Rothermund, K. (2011a). I like it because I said that I like it: Evaluative conditioning effects can be based on stimulus-response learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(4), 466–476. 10.1037/a0023077

Gast, A., & Rothermund, K. (2011b). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition & Emotion*, *25*(1), 89–110. 10.1080/02699931003696380

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

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative–propositional evaluation model. *Social Psychological Bulletin*, *13*(3), e28024. 10.5964/spb.v13i3.28024

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*(5), 573–589. 10.1002/ejsp.166

Gawronski, B., Hu, X., Rydell, R. J., Vervliet, B., & De Houwer, J. (2015). Generalization versus contextualization in automatic evaluation revisited: A meta-analysis of successful and failed replications. *Journal of Experimental Psychology: General*, *144*(4), e50–e64. 10.1037/xge0000079

Gawronski, B., & Quinn, K. A. (2013). Guilty by mere similarity: Assimilative effects of facial resemblance on automatic evaluation. *Journal of Experimental Social Psychology*, *49*(1), 120–125. 10.1016/j.jesp.2012.07.016

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. In James M. Olson (Ed.), *Advances in Experimental Social Psychology* (Vol. 57, pp. 1–52). Elsevier. 10.1016/bs.aesp.2017.06.001

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*(11), e1004567-20. 10.1371/journal.pcbi.1004567

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. 10.1016/j.conb.2010.02.008

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50. 10.1016/j.cobeha.2015.07.007

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

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

Hastie, R., Ostrom, T. M., Ebbesen, E., Wyer, R., Hamilton, D., & Carlston, D. (1980). *Person memory: The cognitive basis of social perception*. Psychology Press.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*(2), 141–154. 10.1016/S0022-1031(77)80007-3

Kerpelman, J. P., & Himmelfarb, S. (1971). Partial reinforcement effects in attitude acquisition and counterconditioning. *Journal of Personality and Social Psychology*, *19*(3), 301–305. 10.1037/h0031447

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. 10.1037/xge0000239

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

Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*. Advance online publication. 10.1016/j.cognition.2021.104792

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*(13), 6035–6044. 10.1073/pnas.1820238116

Kurdi, B., Krosch, A. R., & Ferguson, M. J. (2020). Implicit evaluations of moral agents reflect intent and outcome. *Journal of Experimental Social Psychology*, *90*, 103990–12. 10.1016/j.jesp.2020.103990

Kurdi, B., Mann, T. C., Axt, J. R., & Ferguson, M. J. (2021a). *The prospect of long-term implicit attitude change: Tests of spontaneous recovery and reinstatement*. [Manuscript in preparation]. Department of Psychology, Yale University.

Kurdi, B., Mann, T. C., & Ferguson, M. J. (2021b). Persuading the implicit mind: Changing negative implicit evaluations with an 8-minute podcast. *Social Psychological and Personality Science*. Online first publication. 10.1177/19485506211037140

Kurdi, B., Morris, A., & Cushman, F. A. (2021c). *The role of causal structure in implicit evaluation*. PsyArXiv. 10.31234/osf.io/r7cfa

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. 10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. 10.1037/xge0000179

Lindgren, K. P., Baldwin, S. A., Peterson, K. P., Wiers, R. W., & Teachman, B. A. (2020). Change in implicit alcohol associations over time: Moderation by drinking history and gender. *Addictive Behaviors*, *107*, 106413. 10.1016/j.addbeh.2020.106413

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

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, 68(C), 122–127. 10.1016/j.jesp.2016.06.004

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), 1169–1192. 10.1037/xge0000701

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204–217). Guilford Press.

McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, *94*(5), 792–807. 10.1037/0022-3514.94.5.792

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

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352. 10.1037/0278-7393.10.3.333

Meersmans, T., De Houwer, J., Baeyens, F., Randell, T., & Eelen, P. (2005). Beyond evaluative conditioning? Searching for associative transfer of nonevaluative stimulus properties. *Cognition & Emotion*, *19*(2), 283–306. 10.1080/02699930441000328

Merriman, J., Rovee-Collier, C., & Wilk, A. (1997). Exemplar spacing and infants' memory for category information. *Infant Behavior and Development*, *20*(2), 219–232. 10.1016/s0163-6383(97)90024-2

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234. 10.1037/h0031564

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage*, *28*(4), 757–762. 10.1016/j.neuroimage.2005.03.011

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183–198. 10.1017/s0140525x09000855

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

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology*, *46*(6), 1101–1116. 10.1002/ejsp.2266

Moskowitz, G. B. (1993). Person organization with a memory set: Are spontaneous trait inferences personality characterizations or behaviour labels? *European Journal of Personality*, *7*(3), 195–208. 10.1002/per.2410070305

Moskowitz, G. B. (1996). The mediational effects of attributions and information processing in minority social influence. *British Journal of Social Psychology*, *35*(1), 47–66. 10.1111/j.2044-8309.1996.tb01082.x

Moskowitz, G. B., & Roman, R. J. (1992). Spontaneous trait inferences as self-generated primes: Implications for conscious social judgment. *Journal of Personality and Social Psychology*, *62*(5), 728–738. 10.1037/0022-3514.62.5.728

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*(1), 167–184. 10.1037/0022-3514.77.1.167

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. 10.3758/bf03213230

Newman, L. S. (1991). Why are traits inferred spontaneously? A developmental approach. *Social Cognition*, 9(3), 221–253. 10.1521/soco.1991.9.3.221

Olcaysoy Okten, I., & Moskowitz, G. B. (2020). Easy to make, hard to revise: Updating spontaneous trait inferences in the presence of trait-inconsistent information. *Social Cognition*, *38*(6), 571–625. 10.1521/soco.2020.38.6.571

Olcaysoy Okten, I., Schneid, E. D., & Moskowitz, G. B. (2019). On the updating of spontaneous impressions. *Journal of Personality and Social Psychology*, *117*(1), 1–25. 10.1037/pspa0000156

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

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

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27*(1), 26–47. 10.1016/0022-1031(91)90009-u

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, *19*(3), 249–254. 10.1111/j.1467-9280.2008.02076.x

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, 86(1), 61–79. 10.1037/0033-295x.86.1.61

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition*, *4*(1), 1–17. 10.1521/soco.1986.4.1.1

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

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

Rubinstein, R. S., & Jussim, L. (2019). Stimulus pairing and statement target information have equal effects on stereotype-relevant evaluations of individuals. *Journal of Theoretical Social Psychology*, *111*(1), 256–19. 10.1002/jts5.53

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology*, *75*, 54–70. 10.1016/j.jesp.2017.11.009

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Publishing Group*, *15*(8), 549–562. 10.1038/nrn3776

Rydell, R. J., & Jones, C. R. (2009). Competition between unconditioned stimuli in attitude formation: Negative asymmetry versus spatio-temporal contiguity. *Social Cognition*, *27*(6), 905–916. 10.1521/soco.2009.27.6.905

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

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958. 10.1111/j.1467-9280.2006.01811.x

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, *37*(5), 867–878. 10.1002/ejsp.393

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. 10.1016/s1053-8119(03)00230-1

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Scientific Reports*, *7*(1), 14869. 10.1038/s41598-017-12884-5

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, *110*(50), 20040–20045. 10.1073/pnas.1320322110

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*(7277), 49–53. 10.1038/nature08637

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face? Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology*, 86, 103888. 10.1016/j.jesp.2019.103888

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. 10.1207/s15327957pspr0402_01

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld, & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 29–67). Cambridge University Press.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*(10), 1660–1672. 10.1037/0022-3514.37.10.1660

Teachman, B. A., Clerkin, E. M., Cunningham, W. A., Dreyer-Oren, S., & Werntz, A. (2019). Implicit cognition and psychopathology: Looking back and looking forward. *Annual Review of Clinical Psychology*, *15*(1), 123–148. 10.1146/annurev-clinpsy-050718-095718

Todorov, A. T., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, 66(1), 519–545. 10.1146/annurev-psych-113011-143831

Todorov, A. T., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. 10.1016/j.tics.2008.10.001

Trafimow, D., & Schneider, D. J. (1994). The effects of behavioral, situational, and person information on different attribution judgments. *Journal of Experimental Social Psychology*, *30*(4), 351–369. 10.1006/jesp.1994.1017

Uleman, J. S., Hon, A., Roman, R. J., & Moskowitz, G. B. (1996). On-line evidence for spontaneous trait inferences at encoding. *Personality and Social Psychology Bulletin*, *22*(4), 377–394. 10.1177/0146167296224005

Van Dessel, P., De Houwer, J., & Gast, A. (2015). Approach–avoidance training effects are moderated by awareness of stimulus–action contingencies. *Personality and Social Psychology Bulletin*, *42*(1), 81–93. 10.1177/0146167215615335

Van Dessel, P., De Houwer, J., Roets, A., & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. *Journal of Personality and Social Psychology*, *110*(1), e1–e15. 10.1037/pspa0000039

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. 10.1037//0033-295x.107.1.101

Wimmer, G. E., Li, J. K., Gorgolewski, K. J., & Poldrack, R. A. (2018). Reward learning over weeks versus minutes increases the neural representation of value in the human brain. *Journal of Neuroscience*, 38(35), 7649–7666. 10.1523/jneurosci.0075-18.2018

Wimmer, G. E., & Poldrack, R. A. (2021). Reward learning and working memory: Effects of massed versus spaced training and post-learning delay period. *Memory & Cognition*. Online first publication. 10.3758/s13421-021-01233-7

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*(2), 237–252. 10.1037//0022-3514.47.2.237

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