Psychological Review
2013, Vol. 120, No. 1, 255–280

0033-295X/13/$12.00 DOI: 10.1037/a0030972

The Affective Meanings of Automatic Social Behaviors: Three Mechanisms That Explain Priming

Tobias Schröder and Paul Thagard
University of Waterloo

The priming of concepts has been shown to influence peoples' subsequent actions, often unconsciously. We propose 3 mechanisms (psychological, cultural, and biological) as a unified explanation of such effects. (a) Primed concepts influence holistic representations of situations by parallel constraint satisfaction. (b) The constraints among representations stem from culturally shared affective meanings of concepts acquired in socialization. (c) Patterns of activity in neural populations act as semantic pointers linking symbolic concepts to underlying emotional and sensorimotor representations and thereby causing action. We present 2 computational models of behavioral priming that implement the proposed mechanisms. One is a localist neural network that connects primes with behaviors through central nodes simulating affective meanings. In a series of simulations, where the input is based on empirical data, we show that this model can explain a wide variety of experimental findings related to automatic social behavior. The second, neurocomputational model simulates spiking patterns in populations of biologically realistic neurons. We use this model to demonstrate how the proposed mechanisms can be implemented in the brain. Finally, we discuss how our models integrate previous theoretical accounts of priming phenomena. We also examine the interactions of psychological, cultural, and biological mechanisms in the control of automatic social behavior.

Keywords: priming, affective processes, parallel constraint satisfaction, neural networks, computer simulation

Not all human actions are under the actors' intentional control. Some people start extramarital affairs even though they believe in faithfulness. Others watch a movie although they had promised themselves to exercise that night. Yet others buy more things whenever they go to the grocery store than they had put on the shopping list.

The history of psychology contains many attempts to understand the forces that drive human behavior automatically and without the acting person's conscious awareness (e.g., Bargh & Chartrand, 1999; Freud, 1923/1960; Gibson, 1979; James, 1890/1950; Skinner, 1938). The study of experimental priming effects has been one of the most influential scientific paradigms for understanding the automaticity of social behaviors. The accumulation of evidence that the temporary activation of concepts in people's minds influences their subsequent, seemingly unrelated behaviors has become overwhelming. For example, in classic experiments, subconsciously priming college students with the elderly stereotype made them walk down the hall more slowly (Bargh, Chen, & Burrows, 1996, Experiment 2). Priming the African American stereotype led participants to display more hostility (Bargh et al., 1996, Experiment 3) or to perform poorly in a standardized test (Wheeler, Jarvis, & Petty, 2001), compared to control groups. Activating business versus religion concepts in the subjects' minds caused them to behave less cooperatively in economic dilemma games (Smeesters, Yzerbyt, Corneille, & Warlop, 2009). Numerous examples could be added (for reviews, see Bargh, 2006; Dijksterhuis & Bargh, 2001; Wheeler & DeMarree, 2009).

However, the vast amount of demonstrated experimental effects is not matched by theoretical understanding of the underlying psychological mechanisms. This discrepancy provoked John Bargh, one of the pioneers of the behavioral-priming research program, to ask, "What have we been priming all these years?" (Bargh, 2006, p. 147). He called for leaving the paradigm's "childhood" (p. 147) of generating more and more effects behind, in order to move on to the "second-generation research problem" (p. 147) of uncovering the mechanisms by which conceptual primes get transformed into coordinated social action.

In response to Bargh's (2006) call, we propose a unified theory of the representational processes that underlie behavioral priming effects, and we argue that this theory integrates important previous explanatory attempts. Taking into account mechanisms at the psychological, cultural, and biological levels of explanation (cf.

This article was published Online First December 10, 2012.

Tobias Schröder and Paul Thagard, Department of Philosophy, University of Waterloo, Waterloo, Ontario, Canada.

Tobias Schröder was awarded a research fellowship by the Deutsche Forschungsgemeinschaft (German Research Foundation; #SCHR 1282/1-1) to support this work. Paul Thagard's work is supported by the Natural Sciences and Engineering Research Council of Canada. We would like to thank Terrence C. Stewart for important advice on developing the neurocomputational model. Chris Eliasmith, David Heise, Aaron Kay, Neil MacKinnon, Wolfgang Scholl, Dirk Smeesters, Frank van Overwalle, and participants of the social psychology research seminar at the University of Waterloo provided valuable feedback about theoretical ideas and earlier versions of the article.

Correspondence concerning this article should be addressed to Tobias Schröder, University of Waterloo, Department of Philosophy, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1. E-mail: mail@tschroeder.eu

Cacioppo, Berntson, Sheridan, & McClintock, 2000; Thagard, 2012c, in press), we propose three mechanisms that, taken together, can account for the major discoveries related to automatic social behavior:

1. We think that priming can be explained by *parallel constraint satisfaction*, the principle that the mind amalgamates multiple active representations into a single, coherent *Gestalt* (Read & Simon, 2012; D. Simon & Holyoak, 2002; Thagard, 2000). Automatic actions result from the integration of all the currently active representations, including the primed concept, knowledge about the self and the target person of the action, and other aspects of the situation.

2. We suggest that the constraints on automatic behaviors stem from structures of *affective meaning* that are shared across members of one culture (Heise, 2010; Osgood, May, & Miron, 1975). Priming procedures bias experimental participants' conceptual framings of social situations to correspond to specific affective meanings. By automatically aligning their subsequent actions with those meanings, people implicitly reproduce the social order of their culture (Heise, 2007; MacKinnon, 1994).

3. We further argue that brains use *semantic pointers* to generate actions from affective meanings. Semantic pointers are patterns of spiking activity in populations of neurons that represent symbolic concepts (Eliasmith, in press). They possess "shallow meanings" by virtue of their relation to other concepts and to objects in the world, but they have also "deep meanings," as they are composed of and can be expanded into low-level emotional and sensorimotor representations. We think that the decompression of shallow into deep meanings by spiking neurons, which Eliasmith (in press) described in rigorous mathematical terms, provides a biologically plausible explanation for the effects of symbolic concept activation on action which have been described in the behavioral priming literature.

We describe two complementary computational models of priming effects. One is a localist connectionist network, where the interconnected units represent symbolic concepts, which we used in a series of simulations to demonstrate how the above three principles can account for important instances of behavioral priming. The second model is a biologically realistic spiking neuron model that we used in a simulation to elucidate how those mechanisms might be implemented through computational processes in the brain (Eliasmith & Anderson, 2003).

This article is organized as follows. First, we briefly review previous theoretical proposals for explaining priming effects. Then, we elaborate on the three mechanisms we consider as crucial for a comprehensive and integrative theory of automatic social behavior. Next, we describe how we could simulate results from key experiments with a localist neural network model and a related spiking neuron model, both implementing our theoretical ideas. Finally, we discuss our account in relation to the previous theoretical accounts of priming phenomena as well as in relation to its potential to integrate across psychological, cultural, and biological explanations of social behavior.

Previous Theoretical Accounts for the Automaticity of Social Behavior

Bargh et al. (1996) argued that primed social concepts such as traits or stereotypes activate semantically related mental representations of behaviors. This account can be traced to the famous principle of *ideo-motor action* of William James (1890/1950; see also Prinz, 1987; Stock & Stock, 2004). The mere cognitive representation of an action renders carrying out the action more likely. This does not necessarily require any motivational processes or deliberate decision.

According to this view, priming effects occur as a result of spreading activation: For example, priming of a stereotype activates traits included in that stereotype. The trait representations, in turn, activate related behavior representations. Active behavior representations, in turn, cause related movement of the muscles. One of the problems with this explanation of behavioral priming is that the same prime can have very different consequences, from simple activation of memories to the pursuit of abstract goals (see Bargh, 2006, for review). We think it is necessary to specify what exactly is meant by "semantic similarity" between primes and behaviors, and what the precise mechanisms are that underlie the causation of behavior through activating symbolic concepts.

A second theoretical position is that primes activate *interaction goals,* which, in turn, guide social action. As Cesario, Plaks, and Higgins (2006) argued, priming of a stereotype prepares the primed person to interact with a member of the stereotyped group. Accordingly, in their variant of Bargh et al.'s (1996) classic experiment, they observed a higher degree of hostility in the participants' behavior after they were primed with a homosexuality stereotype rather than a racial stereotype like in Bargh's original study. They argued that this can be explained only by motivational processes, not by direct expression of the stereotype, since most peoples' stereotypes of homosexuals contains traits of femininity and weakness rather than hostility and aggressiveness. Others have pointed out that priming representations of other persons activates the relationship goals people associate with these persons (Fitzsimons & Bargh, 2003; Shah, 2003). For example, priming the concept of friend increases the likelihood of someone helping a stranger. The notion of goal activation as the underlying mechanism of behavioral priming has received considerable empirical support. However, there is a problem quite similar to that of the ideo-motor account (same prime, different effects), as it is not always clear beforehand, precisely which or whose interaction goal is primed by a specific procedure (cf. Bargh, 2006, p. 152). Does presentation of the African American stereotype activate the supposed goals of African Americans, or one's own goals toward African Americans (Cesario et al., 2006)? When representations of significant others are evoked, do the other's goals become active (Shah, 2003) or one's goals toward that person (Fitzsimons & Bargh, 2003)?

A third possible account for prime-to-behavior effects is Wheeler, DeMarree, and Petty's (2007) *active-self* model, according to which primes cause changes in the currently activated working self (cf. Markus & Wurf, 1987) and thereby influence subsequent behavior. The self is a complex representational struc-

ture that integrates various types of self-relevant knowledge and is central in governing social interaction (e.g., MacKinnon & Heise, 2010; Thagard, in press). Not all the (explicit or implicit) information individuals have about themselves is available at all times; hence, only a subset of identity-relevant material is active in a specific situation. According to Wheeler et al. (2007), a priming procedure might influence the situational choice of self-information that is currently active. However, the exact mechanism of integrating a primed concept with self-knowledge is not very clear (Wheeler et al., 2007). Does the self-knowledge people have present a limit to the primed concepts their working self would accommodate? Or can the working-self be extended easily with information that was never before contained in the individual's self-knowledge?

The integration of primed concepts with currently active representations is also a core assumption of Loersch and Payne's (2011) situated inference model, another theoretical proposal for understanding priming. Loersch and Payne argued that the effects of priming result from misattributing the concepts activated by a priming procedure to objects of the environment that are currently in the focus of the person's attention. Their model is broader than the active-self account, since it posits that any object, not just the self, may be the source of misattribution. Accordingly, Loersch and Payne claimed that any form of priming, not just behavioral, results from the same misattribution mechanism, depending on the affordances (cf. Gibson, 1979) of the situation. Like the active-self model, the situated inference model does not rigorously specify the mental mechanism by which primed concepts are integrated with other representations of the situation.

A fifth theoretical position is that primes activate complex metaphorical structures that temporarily alter the way the primed person makes sense of the world: "Perhaps, then, what we have been priming these years is a role, a conceptual structure that contains . . . the *perspective* a person in that role would have on the world—the purposes and goals and values that person . . . would have" (Bargh, 2006, p. 155, emphasis in original). This account emphasizes the role of language for organizing automatic social behavior, citing (among others) Lakoff and Johnson's (2003) theory of conceptual metaphor. According to the latter, social concepts as they are used in priming experiments are embedded in a hierarchical structure of ever more complex metaphors that are ultimately grounded in bodily experience. For example, the meaning of love and affection is often understood in terms of "warmth,"[1] resulting from infants' experience of physical warmth whenever they are hugged by their parents (Lakoff & Johnson, 2003). A limitation of this metaphorical-structure view of priming phenomena (as with the theory of conceptual metaphor itself) is its vagueness. Promising as it is in principle, we think it is necessary to spell out the details of the underlying psychological and neural mechanisms in a more rigorous way.

Representational Processes Underlying Prime-to-Behavior Effects: Three Mechanisms

We think we can integrate these previous explanations of priming phenomena with a description of three general mechanisms of information processing and action control that operate at the psychological, cultural, and biological levels of explanation. Mechanisms are systems of parts whose interactions produce regular changes (Bechtel, 2008).

At the psychological level, the parts are representations, the interactions arise from the constraints among representations and the algorithms for maximizing constraint satisfaction, and the resulting change is adoption of those representations that account for all the constraints in the best possible way. In general, priming is a kind of parallel constraint satisfaction, with primes providing new inputs that produce outputs as the result of constraint satisfaction. The different theoretical accounts of priming reviewed in the preceding section have focused each on a different kind of representation, such as the prime, the self, or the target person of an interaction. Parallel constraint satisfaction allows for taking all of them into account simultaneously.

We suggest that the constraints that are crucial for behavioral priming effects come from culture in the form of affective meanings of concepts. The brain can process affective meanings rapidly and generate behavioral responses without the prior formation of conscious intentions. At the same time, affective meanings are organized in semantic structures shared among the members of one culture (Heise, 2010). The culture-based mechanism of behavioral priming is therefore affective meaning maintenance, where the parts are people, the interactions are social communication, and the resulting change is the adoption of shared cultural meanings that furnish the constraints used in individuals' generation of social action. As we show, affective meanings provide a parsimonious way of specifying the complex conceptual structures thought to be guiding behavioral priming effects (Bargh, 2006).

Semantic pointers provide the biological mechanism for affective meaning and parallel constraint satisfaction, where the parts are firing patterns in neural populations, the interactions result from neural operations including excitation, inhibition, and binding, and the resulting changes are the production and operations of semantic pointers controlling the motor system. The semantic pointer mechanism explains the neural processes at work when symbolic representations translate into movements of the muscles. It is thus a computational specification of the ideo-motor principle at the core of all explanations of priming (e.g., Bargh et al., 1996).

In the present section, we elaborate on these three mechanisms. After that, we turn to a description of the two complementary computational models that implement and specify them. In simulations, we demonstrate how these models predict the results of important experiments under the behavioral priming paradigm.

Parallel Constraint Satisfaction

Mental representations can be construed as networks of constraints. Positive constraints exist between elements that go together (e.g., writing an article and advancing one's career), and negative constraints exist between elements that are incompatible (e.g., writing an article and watching a movie). Parallel constraint satisfaction is the mechanism that organizes all the elements of a representation into one holistic, meaningful pattern, where a coherent set of elements remains active while the incompatible

[1] This metaphor even comes up in one of the most influential contemporary models of social perception, the warmth(sic!)-competence model (Fiske et al., 2007). Inspired by this model, Williams and Bargh (2008) demonstrated effects of priming physical warmth on person perception.

elements get rejected (e.g., one decides to watch the movie and no longer thinks about one's career; Thagard, 2000). Parallel constraint satisfaction can be seen as a development of the classic psychological consistency ideas embedded in theories of balance (Heider, 1946), congruence (Osgood & Tannenbaum, 1955), or cognitive dissonance (Festinger, 1957), since their explanatory target is how the mind arranges multiple representations into a coherent Gestalt (Read & Simon, 2012; Shultz & Lepper, 1998; Simon & Holyoak, 2002).

Different algorithms permit the efficient solution of parallel-constraint-satisfaction problems (Thagard & Verbeurgt, 1998), but models of psychological phenomena typically have used connectionist networks (see Bechtel & Abrahamson, 2002; Rumelhart, McClelland, & the PDP Research Group, 1986). They consist of nodes modeling the elements of representations and of connections between these nodes, which may be positive or negative and vary in strength, depending on whether the connected elements are coherent with each other or mutually exclusive. Positive constraints between elements are modeled as excitation: Whenever a node is activated, this activation spreads to the connected nodes. Negative constraints are modeled as inhibition: Whenever a node is activated, this activation prevents activation of the connected nodes. Solving a constraint-satisfaction problem with such a network involves multiple rounds of updating the activation of all the nodes in parallel by summing up the excitatory and inhibitory inputs they receive from all the connected nodes (see the Appendix for mathematical details). Typically, this procedure yields a stable pattern of activation of some elements and inhibition of the others after a limited number of updates. This pattern then can be interpreted as a coherent mental representation at a given point in time (Thagard, 2000).

Parallel-constraint-satisfaction models originated as explanations for cognitive functions such as letter perception (McClelland & Rumelhart, 1981), discourse comprehension (Kintsch, 1988), and analogical mapping (Holyoak & Thagard, 1989). They also have been applied to social psychological phenomena (for reviews, see Read & Miller, 1998; Read & Simon, 2012; E. R. Smith, 1996). One example is Kunda & Thagard's (1996) theory of impression formation, according to which all the different pieces of information one could know about a given person (like stereotypes, traits, or observed behaviors) amalgamate into a coherent impression, as the different elements constrain each other's meaning. For instance, interpretations of elbowing another person can depend on racial stereotypes: When performed by an African American person, an observer may construe this behavior as violent, in order to fit with the aggressiveness trait that is part of some people's African American stereotype (Kunda & Thagard, 1996, p. 285). However, the same behavior could mean something different to the same observer if it were performed by a Caucasian person, whose stereotype lacks the hostility trait. In this case, people may tend to interpret their observation as a jovial shove rather than a violent attack (Kunda & Thagard, 1996, p. 285). Overall impressions thus result from accounting for all the available information automatically and in parallel.

We propose that the same mechanism of satisfying multiple constraints in parallel guides the operation of behavioral priming effects. Different authors have suggested different kinds of concepts, such as self versus target versus environment representations, to be shaped by priming procedures and thereby causing the behavioral responses observed in so many experiments (for review, see above, or Wheeler & DeMarree, 2009). We think that they all may play a role by putting constraints on the behavior, and we suggest that satisfying them in parallel yields a holistic representation that, in turn, calls for a semantically related action.

This explanation is consistent with classic theorizing about the role of spreading activation for semantic processing, preceding the work on behavioral priming (e.g., Collins & Loftus, 1975; Quillian, 1967; see Bargh et al., 1996). The main difference between spreading activation and parallel constraint satisfaction models is that the latter also include inhibition. Inhibition, in turn, has played a role in studies of negative priming, where inhibiting attention on a stimulus impairs but does not facilitate subsequent retrieval of semantically similar stimuli from memory (Tipper, 1985; for review, see Tipper & Weaver, 2008). We conclude that parallel constraint satisfaction, which combines spreading activation and inhibition, explains priming in general. Below, we demonstrate the plausibility of this claim for the domain of behavioral priming.

With Bargh (2006), we believe that the source of the constraints on automatic behaviors are the conceptual structures that organize people's minds and that are synchronized among members of one culture who share a common language. However, we propose to be more specific about conceptual structures by focusing on the affective meanings of concepts (Osgood et al., 1975). Previous research, to be reviewed in the following section, has shown that affective meanings provide a simple, computationally efficient, yet effective mechanism for aligning social behaviors with cultural meaning structures. We believe that this mechanism can explain behavioral priming, when combined with parallel constraint satisfaction. Primes carry specific affective meanings, and in order to cause an action they need to be integrated with the affective meanings of other representations already active in the primed persons' minds. The first of the computational models described below specifies how we think the mechanism works.

Affective Meaning Maintenance

Affect control theory states that people choose their social behaviors so that the impressions resulting from these behaviors match the affective meaning of the situation (Heise, 1979, 2007; MacKinnon, 1994; Smith-Lovin & Heise, 1988). This mechanism allows for the interpersonal coordination of social action, since it was shown that members of one culture largely agree on the affective meaning of social roles, institutions, settings, and behaviors (Heise, 2010). Following their motive of affective meaning maintenance, social interactants are thus assumed to automatically reproduce the "expressive order" of their culture (Goffman, 1967; Heise, 2007).

For example, we expect mothers to hug babies, rather than to hurt them. The reason for this expectation, following the proposition of affect control theory, is that the affective meanings of the concepts mother and baby fit with the affective meaning of hugging, but contradict the affective meaning of hurting. Creating affective consistency of actors, actions, and targets of actions is thus the proposed mechanism by which cultural norms become effective in people's mental representations of social situations. In principle, the same mechanism applies to the control of action itself, as people apply linguistic labels to their representations of themselves. A person who defines herself as a mother thereby

assumes the cultural definition of motherhood and incorporates the associated affective meaning in her currently active self-sentiment (MacKinnon & Heise, 2010). This will lead her mind to activate emotionally coherent action tendencies, causing the alignment of her behavior with implicit cultural standards.

Affect control theory is rooted in the interpretative symbolic-interactionist tradition of sociology (Blumer, 1969; Goffman, 1967; Mead, 1934; cf. MacKinnon, 1994), but its methodological approach is rigorously quantitative and aimed at precise mathematical formalization of its core principles and predictions. Building upon Osgood's seminal work related to quantifying the connotative meaning of concepts, scholars in the tradition of affect control theory have compiled large culture-specific sentiment repositories (Osgood et al., 1975; Osgood, Suci, & Tannenbaum, 1957; for review, see Heise, 2010). These are data sets of hundreds and often thousands of linguistic concepts along with average ratings of these concepts by cultural informants on three dimensions usually considered as basic and ubiquitous for the representation of affect (Fontaine, Scherer, Roesch, & Ellsworth, 2007; Morgan & Heise, 1988; Osgood et al., 1957, 1975). First, evaluation (E) refers to good and nice vs. bad and awful feelings and associated action tendencies of approach vs. avoidance. Second, potency (P) denotes the opposite of strength and control as opposed to weakness and ineffectiveness in the appraisal of situations or persons. Third, the activity dimension (A) reflects arousal, and distinguishes between excited and calm affective states as well as active and passive behaviors.

Affect control theory's mathematical model of meaning maintenance (see Heise, 2007, Part II) uses such sentiment data sets as an empirical base for predicting the likelihood of specific actions, given assumptions about currently active linguistic representations of the situation in the minds of the interactants. Systems of nonlinear regression equations based on empirical studies prescribe how evaluation-potency-activity profiles of specific actor and target identities combine into profiles corresponding to behavioral expectations, operationalized in terms of affective similarity to the verbs contained in the sentiment repository (e.g., Schröder, 2011; Smith-Lovin, 1987b). The resulting predictions have been shown to be empirically sound in a variety of studies that compared the likelihood of specific actions computed with the affect control model with observed frequencies of these actions (e.g., Heise & Lerner, 2006; Schröder, Netzel, Schermuly, & Scholl, in press; Schröder & Scholl, 2009).

If Heise and his colleagues are correct that affective meaning maintenance creates alignment of automatic social behavior with cultural expectations bound to linguistic concepts, then such maintenance might be one of the key mechanisms underlying behavioral priming effects. We propose that experimental priming procedures activate affective representations that have systematic semantic links with behavioral tendencies shared within cultures. These links are a result of social synchronization of conceptual structures and behavioral experiences among members of one culture, acquired through both a history of immediate interaction experiences as well as learning a common language.

We believe that there are a number of straightforward commonalities between affect control theory and previous theoretical accounts of priming, reviewed above. Affect control theorists assume that situational cues guide the choices of concepts that people make to interpret social situations (Heise, 2007; Smith-Lovin & Heise, 1988). This view is compatible with all current explanations of behavioral priming. MacKinnon and Heise (2010) proposed that situation-specific representations of a social actor's identity are constrained both by the actor's more permanent self sentiment and institutional aspects of the situation. While they use sociological terminology, MacKinnon and Heise's view clearly parallels Wheeler et al.'s (2007) cognitive active-self account of priming phenomena. According to affect control theory, behavior is not only aligned with the meaning of self but also with the meaning of the target person of the action. This fits with the more relationship-focused explanations of priming (Cesario et al., 2006; Fitzsimons & Bargh, 2003). Finally, the culture-specific sentiment repositories compiled by affect control theorists can be considered as a parsimonious empirical operationalization of the complex conceptual structures that Bargh (2006) hypothesized to guide behavioral priming effects. While distributions of concepts over the evaluation-potency-activity space are certainly a simplification of cultural semantic structures, we believe it is a powerful one and sufficient to explain important examples of priming.

Affect control theory has been treated in a separate literature, but shares many commonalities with classic psychological consistency theories (Heise, 1979; Schröder et al., in press). Thus, its integration with parallel constraint satisfaction, our first proposed mechanism of priming, is straightforward. When generating behaviors, people are supposed to satisfy multiple affective representations (of self, others, situations) in parallel. Accordingly, the first computational model of priming effects we propose below can be understood as a connectionist version of affect control theory.

Most connectionist models of social psychological phenomena are not very biologically realistic, however. The brain does not represent (affective meanings of) stereotypes, traits, or social behaviors as single nodes but, rather, as highly distributed patterns of spiking activity in thousands or millions of neurons. We need to turn to a third mechanism, semantic pointers, in order to be able to understand the biological processes that underlie the control of automatic social behaviors.

Semantic Pointers

According to a biologically realistic architecture of cognition proposed by Eliasmith (in press), semantic pointers are patterns of activity in spiking neurons that carry semantic content by pointing to other neural activity patterns that represent symbols, states of the world, or sensorimotor experiences. The semantic pointer architecture shares many features with classic connectionist ideas of the mind (reviewed above); however, it comes with important advances, aimed at providing a better explanation of how semantic representations are connected with physiological processes and how they enable the organism to interact with the world (cf. Eliasmith, 2005; Parisien & Thagard, 2008). First, the nodes in the network exhibit the biological properties of real neurons, with flows of current being simulated that result in nonlinear spikes, i.e., sudden voltage increases that serve as signals for subsequent, connected neurons (for mathematical details, see Eliasmith & Anderson, 2003). Second, symbolic concepts are represented as patterns of spiking in thousands of artificial neurons rather than as single nodes in the network. Third, mathematical operations embedded in connection weights between different populations of neurons allow the simulation of binding and, hence, the imple-

mentation of compositional and hierarchical representations of concepts (Eliasmith, 2004; Stewart & Eliasmith, 2012). As we make clear in the subsequent paragraphs, the latter aspect is especially important for understanding behavioral priming effects, since it provides an elegant and well-specified mechanism of how abstract conceptual representations may translate into behaviors (i.e., physical actions in a physical world).

Most crucial is the distinction between shallow and deep meanings (Eliasmith, in press). Shallow meanings come from symbol-like relations of activity patterns to objects in the world and to other activity patterns representing symbols. Deep meanings are multimodal and involve relations to perceptual, motor, and emotional information. Figure 1 uses the concept of "mother" to illustrate the relationship between shallow and deep meanings in biologically plausible mental representation. The concept is represented by a unique pattern of activity in a population of neurons (symbolized by the group of small circles in the top row of Figure 1), which results from binding connected activity patterns. These may be of symbolic nature, such as the conceptual definition that mothers are female family members (of course, this is an incomplete definition, but we would like to keep the example as simple as possible). In this case, "female" and "family" are represented by other, distinct activity patterns. So far, this is not more than a neural implementation of the semiotic premise that the meaning of a symbol is given by its relation to other symbols (Peirce & Welby-Gregory, 1977). However, symbols need to be "grounded," in order to enable the organism to use them as tools for interacting with the real world (Harnad, 1990, see also Parisien & Thagard, 2008). According to the semantic pointer hypothesis, such grounding is given by deep meanings. In the example, the meaning of mother is also constrained by patterns of neural activity which represent (a) emotions such as love and happiness; (b) a memory of past interactions with one's mother such as being hugged or fed when one was an infant; and (c) the sensations going along with these experiences such as feeling the physical warmth accompanying any contact between the bodies of mother and child (cf. Lakoff & Johnson, 2003).

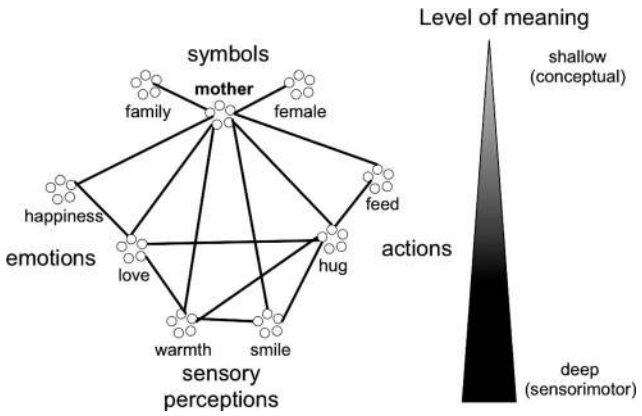

Figure 1. Representation of social concepts in a semantic pointer. Small circles indicate firing patterns in populations of neurons. Solid lines denote binding/compression. In this example, the meaning of "mother" arises from the constraints imposed by related symbolic concepts such as family or female (shallow meaning) but also from incorporating underlying representations of emotional, sensory, and action experience with "mothers" (deep meaning).

The key mechanism of semantic pointers that enables brains to represent complex conceptual structures is binding of symbolic, affective, and sensorimotor representations into a single activity pattern (Eliasmith, 2004, in press; Stewart & Eliasmith, 2012). The semantic pointer hypothesis is thus compatible with the view of Damasio (1989) that the brain binds entities and events by multiregional activation from convergence zones but is far more specific about how this binding works. Binding by semantic pointers is relevant to explaining the nature of concepts, intentions, emotions, and creativity (Blouw, Solodkin, Eliasmith, & Thagard, 2012; Schröder, Stewart, & Thagard, 2012; Thagard, 2012b; Thagard & Schröder, 2012).

Most importantly for understanding behavioral priming phenomena, the binding can be reversed. Primes create conceptual representations that link affect, memories, and behavior. These semantic pointers can be "decompressed" into the underlying deep meanings, by evoking related sensory and motor representations (for mathematical details, see Eliasmith, in press; for empirical evidence, see Andres, Olivier, & Badets, 2008; Barsalou, 1999; Glenberg & Kaschak, 2002; Springer & Prinz, 2010).

The semantic pointer hypothesis is a computational specification of theories that human cognitions, emotions, and conceptual representations are partially embodied in perceptual and motor experience (e.g., Barsalou, 1999; Crawford, 2009; Lakoff & Johnson, 2003; Niedenthal, Winkielman, Mondillon, & Vermeulen, 2009; Williams & Bargh, 2008; Williams, Huang, & Bargh, 2009). By connecting the shallow meanings of symbols with the deep meanings of emotion and action through the spiking activity of neurons, it also provides biological plausibility to the mechanism of affective meaning maintenance, reviewed above. We think that Osgood's et al. (1957) semantic differential and Heise's (2007) affect control model of action regulation reflect the relation between the shallow and deep meanings of social interaction.

Summary

We propose the following multilevel explanation of behavioral priming effects, which synthesizes psychological, cultural, and biological mechanisms. Priming procedures activate concepts in the minds of the primed persons. The psychological mechanism of parallel constraint satisfaction explains how these concepts get combined with other representations currently active, most notably those of the self and potential interaction partners, and potentially other features of the environment. The result is a holistic representation of the self in a situation, which implies specific action tendencies. The constraints of concepts on behaviors are given by affective meanings, shared among members of one culture as a result of previous interactions and language learning during socialization. At the cultural level, the mechanism is thus affective meaning maintenance, creating aligment of automatic social actions with the expressive order of society. The operations of semantic pointers constitute the biological mechanism underlying priming. The brain represents affective meanings through multimodal patterns of spiking activity in distributed neural populations. Shallow meanings at the conceptual level can be decompressed into deep meanings at the sensorimotor level, which is how the

activation of concepts can cause physical actions in a physical world.

Models and Simulations

In order to substantiate and specify our theory of priming, we have developed two complementary computational models and used them to simulate the results of important experiments from the literature on automatic social behavior. We use these models' ability to generate data patterns that match existing experimental data as evidence for the power of our proposed three mechanisms to explain behavioral priming (cf. Thagard, 2012a, pp. 8–11). The first model is a localist parallel-constraint-satisfaction model similar to the model of impression-formation by Kunda and Thagard (1996). These simple models have the advantage that they can be used flexibly to simulate a variety of complex phenomena, explaining what happened in many priming experiments. The disadvantage of localist models is that they provide only a very rough approximation to the underlying biological processes. Accordingly, we describe a second model, based on the Neural Engineering Framework of Eliasmith and Anderson (2003). The advantage of greater biological plausibility comes at the price of a tremendous increase in complexity with spiking patterns in thousands of neurons with biologically realistic parameters. Consequently, we provide only one example simulation with the second model, intended as a proof-of-principle that the mechanisms we proposed in the previous section can indeed be computed by a real brain.

Model 1: Parallel Constraint Satisfaction

This model builds upon Kunda and Thagard's (1996) model of impression formation, both theoretically and technically. Theoretically, we construe priming effects as a result of parallel processing of all the information available in the specific situation. Technically, we used a modified version of IMP (for IMPression formation), the computer program written in LISP to implement the parallel-constraint-satisfaction theory of person perception (see the Appendix of Kunda & Thagard, 1996).

Our connectionist model of behavioral priming comes with three important advances over the earlier IMP model. First, it includes mental models of the self and action beyond the representation of another person. By including separate representations for the self, target person, and behavior, we sought to align the model theoretically not only with affect control theory but also with the active-self, relationship-oriented, and direct expression accounts of behavioral priming. Second, the model has central nodes for affective meanings, which we propose as a connectionist implementation of Heise's (1979, 2007) idea that representations of the self, other persons, and behaviors can be operationalized as patterns of evaluation, potency, and activity. Third, the connection weights between conceptual nodes are entirely determined by empirical data, taken from previous rating studies under the affect control theory research program. This strategy models and quantifies the conceptual structures that Bargh (2006) proposed to guide behavioral priming effects. We show below that the very same, theoretically driven model structure, combined with parameter adjustments solely determined by independent empirical data, results in accurate simulations of a wide variety of experiments in the behavioral priming literature.

The model is displayed in Figure 2. The center represents situational affective representations of the self, the target person of the action, and the behavior as interconnected patterns of evaluation (E), potency (P), and activity (A). This is a connectionist implementation of Heise's (1979, 2007) suggestion that the control of social behavior is governed by a desire to maintain affective meanings.[2] In Figure 2, solid lines are excitatory connections between nodes. The more positively one perceives oneself (Self-E) and one's interaction partner (Target-E), the likelier it is that one will exhibit positive behaviors (Behavior-E). These connections reflect the well-established psychological principle of evaluative balance going back to Heider (1946). For the potency dimension, we assume an excitatory connection between self and behavior (the more powerful one feels, the more dominant one acts), but an inhibitory link between the representation of the target person and one's action. This reflects the well-known principle of complementarity from Interpersonal Theory, according to which people tend to respond submissively to displays of power and vice versa (Carson, 1969; Sadler, Ethier, & Woody, 2011; Tiedens & Fragale, 2003). For activity, we assume positive constraints between self, target, and behavior representations, in line with work on emotional contagion which has shown that (verbal and nonverbal) action is an important vehicle for the transmission of affective states from one person to another (e.g., Hatfield, Cacioppo, & Rapson, 1994).

The network input consists of symbolic representations of the self, the target person, and the primed concept (see the lower part of Figure 2). These are all activated by the IMP program's "observed" function (see Kunda & Thagard, 1996, or the present Appendix for details). The prime is allowed to influence behavior, self, and target representations, in order to account for the different pathways brought up in the behavioral priming literature (for review, see Wheeler & DeMarree, 2009). However, in determining the overall state of the network, the prime competes with other currently active knowledge about the self and the other person (depicted as "self proxy" and "other proxy" in Figure 2). In the simulations of behavioral priming experiments described below, the symbolic input nodes correspond to the independent variables used in the experiments. The connection weights between the self, prime, and target nodes on the one hand, and the evaluation, potency, and activity nodes on the other, are determined by multiplying the default values of the IMP program with the mean ratings of the respective concepts from empirical studies aimed at creating cultural repositories of affective meaning (see Heise, 2010, for review).

The network output consists of a symbolic representation of action (the top node in Figure 2), corresponding to the dependent variables in the simulated experiments. The connection weights

[2] Affect control theory's algebraic models of meaning maintenance, based on empirical rating studies in different cultures (e.g., Schröder, 2011; Smith-Lovin, 1987b), suggest additional and more sophisticated relationships between affective representation of self, target, and actions than the ones implemented in our present model. For example, the equations contain significant coefficients for two- and three-way interactions both within and across Osgood's dimensions. For parsimony, we implemented only those constraints in our network that can be linked very clearly and undoubtedly to well-established social psychological theories, and this was enough to produce all the simulations described here.

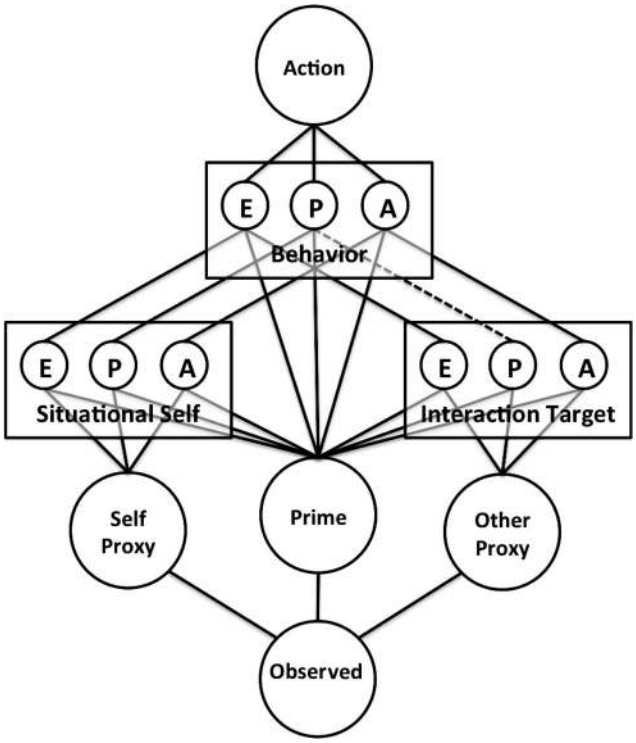

Figure 2. Parallel-constraint-satisfaction model of priming effects. The "observed" node (cf. Kunda & Thagard, 1996) in the bottom row activates the primed concept as well as symbolic representations of self and the target person of the interaction. These, in turn, excite (or inhibit) activation of evaluation (E), potency (P), and activity (A) patterns of self, target, and behavior. The top node is a symbolic action representation (dependent variable in prime-to-behavior experiments) and will be either activated or inhibited once the constraint network has settled. All the connection weights are determined by empirical E-P-A ratings of concepts in previous studies.

between the action node and the evaluation, potency, and activity nodes are determined by empirical data from independent studies.

In sum, the nework structure corresponds to well-known facts from psychology, and the input comes from empirical surveys of cultural attitudes. The resulting output corresponds to experimental data from studies of behavioral priming.

Rationale for simulations. We simulated important examples of priming experiments in order to test our model. For each simulation, we recreated the experimental conditions by adjusting the connection weights of the primed concept according to the average evaluation, potency, and activity rating taken of that concept from existing repositories of cultural sentiments. We simulated variation among participants by running the model 1,000 times for each experimental condition. The number of simulation runs, which by far exceeds the typical number of participants in an experiment, was chosen to ensure the stochastic stability of the model and, hence, the replicability of the simulations. In each model run, the computer set the exact connection weights for the self, prime, and target person nodes by drawing from a random Gaussian distribution centered around the mean empirical evaluation, potency, and activity rating, respectively, and using the empirical standard deviations of the same ratings.[3] This procedure ensured that our model allows for idiosyncratic variations and situational fluctuations of affective meanings, while still corresponding on average to cultural norms provided by informants independent from the specific samples of the original experiments.

The mathematical details of the simulations are given in the Appendix. A rough verbal description of the algorithm goes as follows. The "observed" node activates the self, prime, and target nodes (cf. Kunda & Thagard, 1996). Then, all the other nodes in the network update their activations according to the summed activations they receive from the nodes to which they are connected. This updating occurs in parallel. Then, a next cycle of updating begins, where all nodes recompute their activation by taking changes in activation levels of the connected nodes from the previous cycle into account. This updating continues until the whole network exhibits a pattern of activation that causes no further changes by spreading or inhibiting activation. Usually, it takes between 70 and 200 updating cycles until the network settles in that stable state. Whenever a network had reached a stable state, the computer recorded the activation level of the action node in this stable state.

At the end of each simulation, we thus had 1,000 activation parameters per experimental condition of the dependent variable in question. We treated them with statistical procedures typically used for real experimental data. Metaphorically, our analyses can be conceived of as an examination of data provided by virtual experimental subjects. To test our model, we compared these data patterns with the ones reported in the respective original studies.

The selection of experiments for simulations was guided by the motivation to include examples from each of the different lines of theorizing reviewed in the introduction (priming as direct expression of traits and stereotypes, priming as activation of interaction goals implied by stereotypes and types of relationships, priming as activation of self-knowledge). The selection was constrained by the availability of relevant concepts in the databases of affective meanings that we used as sources of input for the simulations (Francis & Heise, 2006; Schröder, 2011; Schröder, Rogers, Ike, Mell, & Scholl, 2012).

Simulation 1: Trait activation. In Bargh et al.'s (1996, Experiment 1) first demonstration of behavioral priming phenomena, subjects were primed with the concepts of rudeness versus politeness versus a neutral control condition. The dependent variable of the experiment was whether and when they would interrupt the experimenter from talking to another person during a 10-min interval. It was shown that interruption happened more quickly and frequently than in the control condition, when participants had been exposed to rude-related stimuli. It happened less quickly and frequently, when they had been primed with polite-related stimuli.

For our simulation, we used data from a repository of affective meanings that contains average evaluation-potency-activity (EPA) ratings for 1,500 concepts, provided by U.S. undergraduates (Francis & Heise, 2006). On scales ranging from −4 for bad (E–), weak (P–), or passive (A–) through 0 for neutral to +4 for good (E+), strong (P+), or active (A+), the adjective "rude" has the following EPA profile: [–2.69/-0.74/0.60]. The corresponding profile for "polite" is [2.86/1.68/0.41]. As explained above, we used these

[3] We used LISP code by Percival (1993) to implement this randomization.

profiles for adjusting the connection weights of the prime node in the model (see Figure 3).[4] The neutral control condition was simulated with a [0/0/0] profile. Francis and Heise's (2006) data set also contains mean ratings for the verb "interrupt" [–1.51/-0.12/1.15], which we used for adapting the connections of the action node in the model.

Since Bargh and his colleagues examined only main effects and did not report any data regarding self-concepts or target representations, we used the following default EPA profiles for the respective connection weights in our model (the self and target proxy nodes): For self, we took mean EPA ratings of "Myself as I really am" [1.97/0.75/1.04] from a recent study of stereotyping (Schröder, Rogers, et al., 2012). For target, we used ratings of "student" [1.93/0.92/1.20] from Francis and Heise's (2006) data set, since psychological experiments usually happen on campus where fellow students are the most likely interaction partners of experimental participants. Note that we used these self/target proxies in all the following simulations, unless there were specific manipulations in the original studies that required a different input.

Our simulations match the pattern of data reported by Bargh et al. (1996). The average activation of the "interrupt" action node in the simulations that used EPA ratings of "rude" for setting prime connection weights was $M = 0.57$. The corresponding parameters in the "neutral" and "polite" conditions were $M = -.10$ and $M = -.55$, respectively. A one-way analysis of variance (ANOVA) revealed that this parameter difference across conditions was statistically significant, $F(2, 2997) = 2{,}366.00$, $p < .001$. The activation parameter in our model is theoretically related to the accessibility of the related representation, in this case the action of interrupting someone. Therefore, our simulation matches Bargh et al.'s finding that participants primed with rude-related concepts were more likely to interrupt the experimenter.

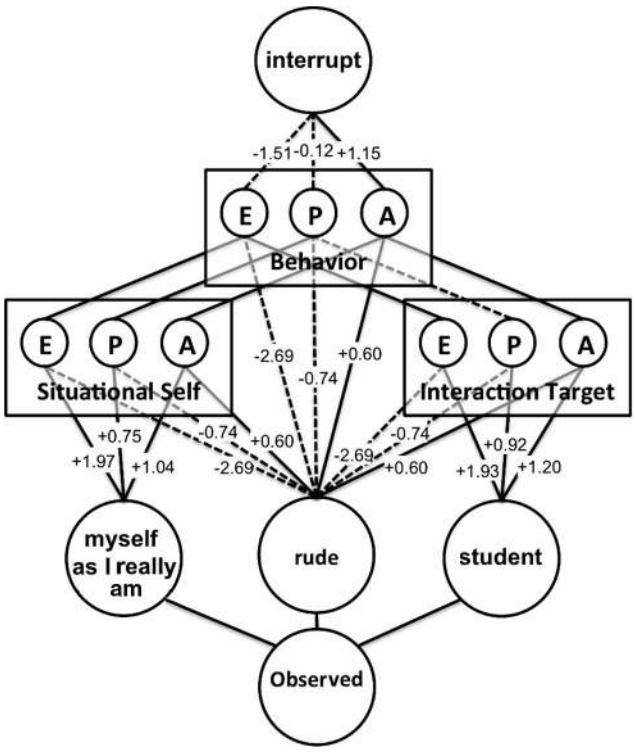

Figure 3. Adaptation of connectionist model from Figure 2 for Simulation 1 (rudeness condition of Bargh, Chen, & Burrows's, 1996, Experiment 1). Connection weights stem from mean EPA (evaluation, potency, activity) ratings of the respective concepts (Francis & Heise, 2006).

A different way of interpreting the simulation outcome is given by the absolute direction of the activation parameter. If we consider each run of the model as a simulation of one participant's information processing during the experiment, we can treat any resulting positive activation value of the "interrupt" node in the network as if that simulated participant decided to interrupt the experimenter, whereas any negative value (= inhibition of the action) corresponded to the decision not to interrupt them. The original data of this from Bargh et al.'s (1996) experiment is displayed along with our simulation results in Figure 4. It can be seen that the simulation reproduced the original findings, including the baseline level of the control condition, while overestimating the effect size. The overestimation can be explained by noting that real interactions are constrained by many additional representations such as the setting, general norms, attractiveness of the target person, etc., which our model does not capture but which are likely to attenuate the effects of a priming procedure. The important point is that our model, constrained only by general theoretical principles and fully independent survey data, can reproduce the pattern of data in this classic priming experiment. Hence, parallel constraint satisfaction, where the constraints are affective meanings, explains this case of behavioral priming.

Simulation 2: Racial stereotypes. Our next simulation deals with the effect of activating an African American stereotype on the perceived hostility of the experimental participants' behavior (Bargh et al., 1996; Experiment 3). The procedure was exactly the same as described under Simulation 1, except that the connection weights of the prime and action nodes in the network were modified. To simulate the African American priming condition, we used the average evaluation-potency-activity (EPA) rating of "Blacks" [0.53/0.19/1.15] for the Caucasian condition the EPA profile for "Whites" [0.80/2.39/0.85], from Schröder, Rogers, et al.'s (2012) study on affective meanings of stereotypes. There was no control condition in the original experiment. The dependent variable was simulated with the ratings of "hostile" [–1.56/0.03/0.49] from Francis and Heise's (2006) repository of U.S. cultural sentiments. The way the network was set up for Simulation 2 is displayed in Figure 5.

The simulation resulted in the following average activation parameters for the "hostile" action node: $M = -0.24$ (African American condition) and $M = -0.30$ (Caucasian condition). The difference is statistically significant, $t(1958.9) = 2.84$, $p < .01$. This fits with data reported by Bargh et al. (1996), according to which participants primed with an African American stereotype were judged to behave in a more hostile way by observers blind to the experimental condition. Note that our simulation predicts a low

[4] The Francis and Heise (2006) data set contains separate data for male and female raters. For the present purpose, we used averages weighted by sample size. As described above, we also used the empirical standard deviations in the random procedure aimed at simulating interindividual variation in affective meanings of the primed concepts. In order to keep our descriptions accessible, we do not report them here, but they can be obtained on request from the first author.

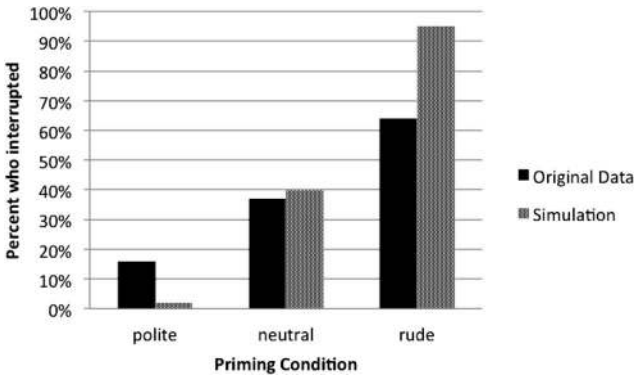

Figure 4. Dark color—Percentage of participants primed with rudeness, politeness, or control concepts who interrupted the experimenter in Bargh, Chen, and Burrows's (1996) Experiment 1. Light color—Results of simulating the experiment with the present parallel-constraint-satisfaction model. Displayed is the percentage of simulations in which the action "interrupt" had any positive amount of activation after the network had settled. Adapted from "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action Permission," by J. A. Bargh, M. Chen, and L. Burrows, 1996, *Journal of Personality and Social Psychology, 71,* p. 235. Copyright 1996 by the American Psychological Association.

overall baseline of hostility. For both priming conditions, the activation parameter is negative, roughly meaning that displaying hostility after priming of African American is *unlikely* and after priming of Caucasian *very unlikely*. This base rate prediction also matches Bargh et al.'s (1996) data, since the corresponding hostility ratings were $M = 2.79$ versus $M = 2.13$ on "a scale of hostility ranging from 0 (*not at all hostile*) to 10 (*extremely hostile*)" (p. 239, italics in original).

Simulation 3: Motivated preparation to interact—the gay stereotype. Cesario et al. (2006) proposed that behavioral effects of priming stereotypes are produced by preparing the primed person to interact with someone from the stereotyped group, rather than through direct expression of the stereotype. To support this argument, they present a modification of the Bargh et al. (1996) experiment discussed above, where they primed stereotypes related to sexual orientation rather than to race. They argue that the stereotype of gay men includes passivity and femininity rather than hostility. Therefore, increased hostility in the participants' behavior after being primed with homosexuality-related concepts could only be explained by the authors' account focusing on motivation, but not by simpler accounts focusing on cognitive activation.

In order to test the claim that our model can integrate such competing theoretical explanations, we simulated the Cesario et al. (2006) variant of stereotyping effects as well. The simulation was exactly the same as the previous one, except that we exchanged the data for "Black" versus "White" primes with EPA profiles for "homosexual" [0.63/−0.04/0.63] vs. "heterosexual" [1.62/1.23/0.71], taken from the Francis and Heise (2006) data set (see Figure 6 for a visualization of the network in this simulation). This worked as we expected: In line with the results reported by Cesario et al., the mean average activation of the "hostile" action node was significantly higher ($M = –.23$) in the homosexual condition than in the heterosexual condition ($M = –.38$), $t(1935.2) = 6.99$, $p < .001$. Again, the fact that for both conditions, the activation parameter was negative, suggesting hardly any hostility at all, matches the very low baseline level of hostility reported in the original study ($M = 1.18$ vs. $M = 0.44$, respectively, on scales ranging from 0 to 10; see Cesario et al., 2006, p. 898).

Simulation 4: Relationship goals. Fitzsimons and Bargh (2003) argued that people associate specific goals with different types of relationships, and they presented empirical data corroborating the notion that people align their behaviors with these goals after being primed with a specific relationship. In their Study 1, they found that travelers waiting at an airport were more willing to commit to helping a stranger after they had answered a series of questions about a "friend," in comparison to control subjects who had answered the same questions about a "coworker."

We simulated this experiment as follows. First, in order to account for the interaction of a traveler with a stranger at an airport, we adjusted the connection weights of the self and target proxy nodes in the model so that they matched Francis and Heise's (2006) mean empirical ratings of "traveler" [1.58/0.49/0.94] and "stranger" [–0.03/-0.17/-0.34], respectively. For the primes in the two experimental conditions, we used EPA profiles for "friend" [3.12/2.19/1.66] and "coworker" [1.04/0.49/0.56], and for the dependent variable ratings for the verb "help someone" [3.01/2.57/1.44], all from the same repository (see Figure 7).

The average activation of the "help someone" action node was significantly higher in the "friend" ($M = 0.79$) than in the "coworker" ($M = 0.54$) condition, $t(1071.1) = –13.74$, $p < .001$, matching Fitzsimons and Bargh's (2003) data. If we interpret any

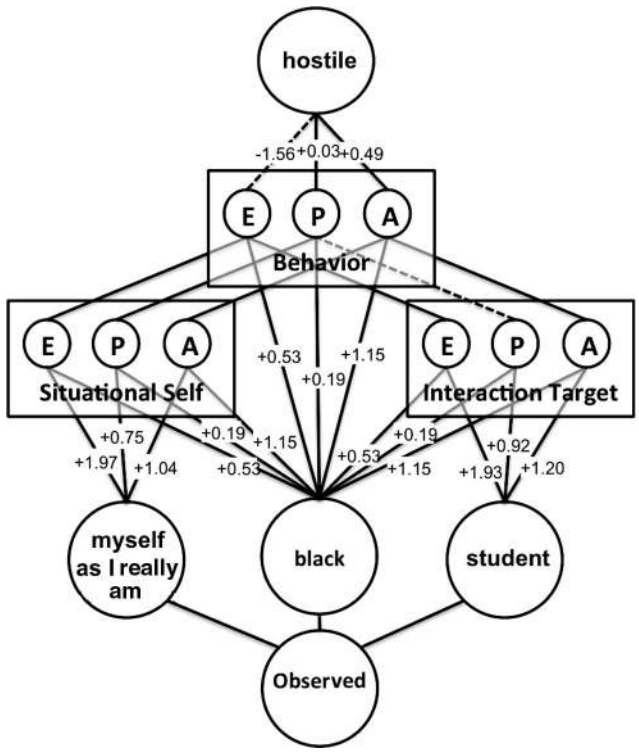

Figure 5. Adaptation of connectionist model from Figure 2 for Simulation 2, African American condition of Bargh, Chen, and Burrows's (1996) Experiment 3. E = evaluation; P = potency; A = activity.

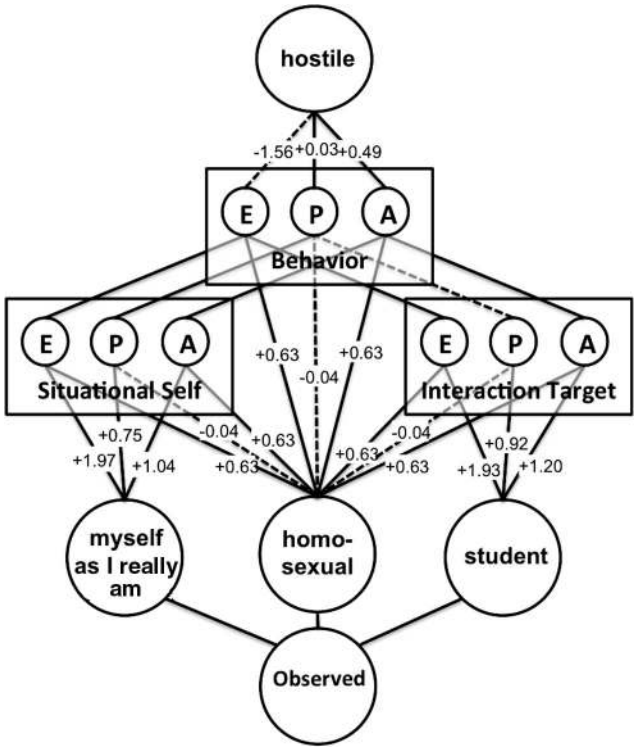

Figure 6. Adaptation of connectionist model from Figure 2 for Simulation 3, homosexuality condition of Cesario, Plaks, and Higgins's (2006) Experiment 1. E = evaluation; P = potency; A = activity.

positive activation parameter as the decision to help the stranger and any negative value as the denial of help, our model suggests that after the "friend" priming, 99.4% of subjects would choose to help, in contrast to 84.6% following the "coworker" prime. This is clearly a much too high base rate in comparison to the data from the original study (nine out of 17 and three out of 16 helped, respectively). However, in our model helping comes at no cost, whereas in the original study, subjects were asked to fill out a lengthy questionnaire after they had already completed one. The lack of representing such external objects is one of the many obvious simplifications of our model. Most importantly for the present purpose, the pattern of helping decisions across the priming conditions is the same in our simulations and in the experiment. The authors tested the data pattern with a nonparametric test, and so did we with the successful simulation result, $\chi^2(1, N = 1000) = 146.80$, $p < .001$.

Simulation 5: Priming effects moderated by self-knowledge. In our final simulation with the connectionist model, we address the active-self account of priming effects (Wheeler et al., 2007), according to which priming procedures influence behavior through the temporary alteration of subjects' self-representation. The main prediction resulting from this perspective is that the self-knowledge currently active poses an important constraint on participants' susceptibility to priming. As an example for this line of work, we take two experiments by Smeesters et al. (2009), but it should be noted that evidence for the general phenomenon is abundant (e.g., DeMarree & Loersch, 2009; DeMarree, Wheeler, & Petty, 2005; Dijksterhuis & Van Knippenberg, 2000; for review, see Wheeler et al., 2007).[5] Smeesters and colleagues examined competitive versus cooperative behavior in an economic game, after priming business- versus religion-related concepts. Subjects primed with concepts like "manager" behaved more competitively than those primed with stimuli like "priest," but this effect was moderated by the degree to which their self-concept of being either high or low on social-value orientation (SVO) was active, a personality variable assumed to be associated with more cooperative or competitive behavior, respectively. A self-concept can be active because of its ongoing accessibility, examined in Smeesters et al.'s Experiment 1, or because it was activated through an experimental procedure, examined in Experiment 2 in the same article. Our model cannot distinguish between these two reasons for a self-concept to be currently active; hence, we present only one simulation for the experiments, both of which produced largely similar data patterns.

We operationalized self-concept accessibility through altering the connection between the self-proxy and the "observed" node (see Figure 8). The previous four simulations assumed self-representation to be active, since the self-proxy always received initial activation directly from the "observed" node. In order to simulate the experimental conditions where self-knowledge was present but not currently active, we simply deleted the direct excitatory pathway between the observed and self nodes. In these cases, whether the particular kind of self-knowledge would become active in the simulation was a matter of the process of parallel constraint satisfaction.

Subjects in Smeesters et al.'s (2009) experiments were from Flanders, the Dutch-speaking part of Belgium. We chose to use a German evaluation-potency-activity dictionary (Schröder, 2011) as database for the simulation because of the lack of a cultural sentiment repository in Dutch and the greater similarity of Dutch culture with German, as opposed to U.S. American, culture (Gupta, Hanges, & Dorfman, 2002; in personal communication, April 27, 2011, Smeesters agreed with this reasoning). EPA profiles for "cooperative" [2.46/1.03/0.08] and "competitive" [–0.71/1.48/1.71] were used to simulate high vs. low social value orientation. Mean ratings for "manager" [–1.24/2.30/1.71] and "priest" [0.93/0.44/-1.30] served as adjustments of the connection weights

[5] Some recent events, which became known publicly after submission of the present article for publication, have cast doubt on the credibility and academic integrity of Dirk Smeesters, the first author of the studies addressed in Simulation 5. He resigned from his position at the Erasmus University of Rotterdam, citing personal reasons, after the University had asked for retraction of three of his more recent studies (two published, one submitted), following an examination of all his work by a special Committee for Inquiry into Scientific Integrity. The retractions were based on statistical irregularities in the published data (data "too good to be true"); Smeesters admitting that, without stating so in the publications, he deleted data from subjects who had failed a manipulation check to obtain significant results; and the fact that the original raw data were not available for examination by the committee. According to the final committee report (Erasmus University, 2012), none of these problems applied to the publication relevant for Simulation 5 (Smeesters et al., 2009). Specifically, the report states that Vincent Yzerbyt (the second author) conducted the data analysis jointly with Dirk Smeesters and that the original raw data were available. We thus have no reason to doubt the credibility of these two studies, the main results of which were reproduced by our model in Simulation 5. In addition, numerous other studies without Smeesters's involvement have demonstrated that activation of the self-concept constrains effects of priming on behavior, so we can consider this phenomenon as established knowledge (see references in the main text).

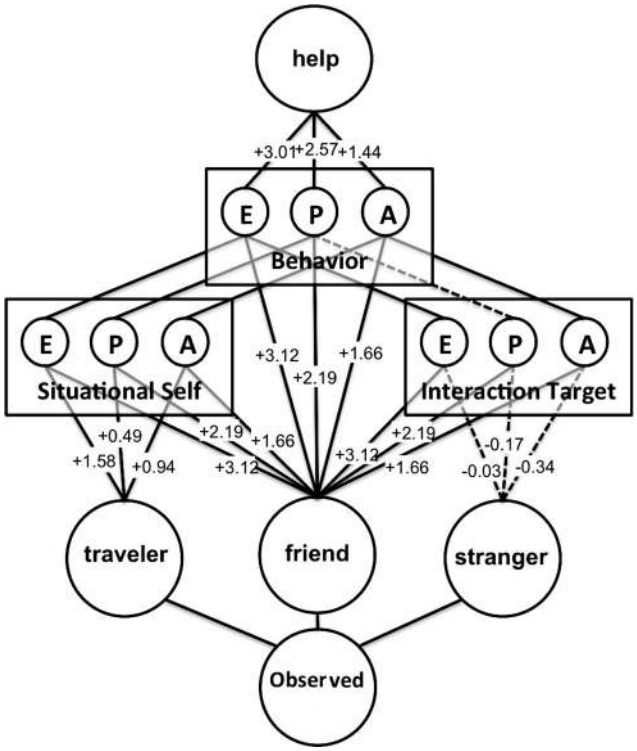

Figure 7. Adaptation of connectionist model from Figure 2 for Simulation 4, friend condition of Fitzsimon and Bargh's (2003) Experiment 1. E = evaluation; P = potency; A = activity.

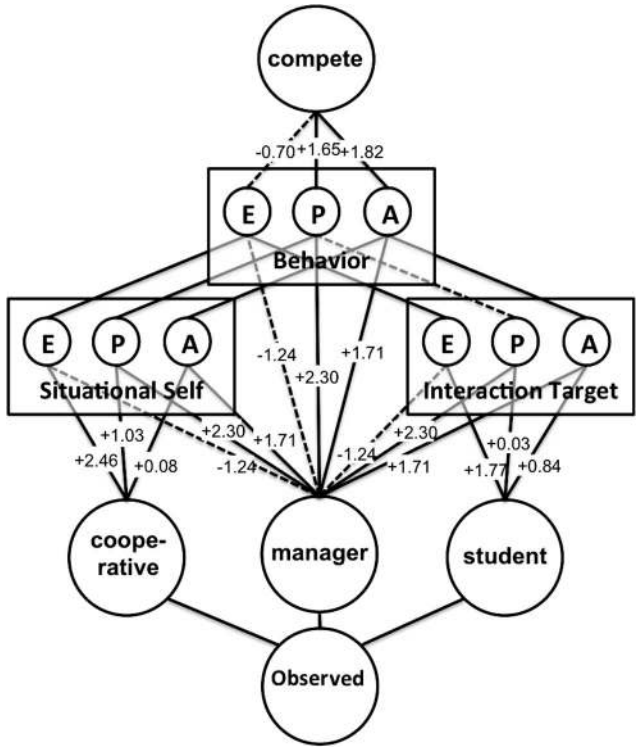

Figure 8. Adaptation of connectionist model from Figure 2 for Simulation 5 (Accessible Self-Knowledge × High Social Value Orientation × Business Prime Condition of Smeesters, Yzerbyt, Corneille, & Warlop's, 2009, Experiments 1 and 2). Note that for the simulations of experimental conditions where self-concept accessibility was low, the link between "observed" and "cooperative" was cut. E = evaluation; P = potency; A = activity.

in the business and religion conditions of the priming procedure, respectively. The German EPA profile for "student," which we used as a proxy for representing the target person like in simulations 1–3, is [1.77/0.03/0.84]. The verb "compete" [−0.70/1.65/1.82] was chosen to simulate the dependent variable (in the experiment, the number of chips the subjects allocated to themselves rather than to their interaction partner). A visualization of the network structure for one of the eight experimental conditions is provided in Figure 8.

The results of the simulation—the average activations of the "compete" node in the different experimental conditions—are displayed in Table 1. In line with the data reported by Smeesters et al. (2009), there are main effects of priming and social value orientation. In the business priming condition, the "compete" node had a positive mean activation; in the religion condition, it was negative (suggesting inhibition). High social-value orientation (SVO), operationalized by our using EPA data for "cooperative" for the self connection weights, led on average to lower activation of the action than low social-value orientation. However, in the condition with high self-concept accessibility, simulated by forcing activation of the self node, the main effect of priming was attenuated by SVO. In the simulation data of the nonaccessibility condition (rightmost two columns of Table 1), such attenuation is hardly visible. Like in Simulation 1, our model overestimated the relative importance of the priming procedure—in the original experiments, the SVO main effect was much more important in size than the priming effect, contrary to our simulation. As we had argued above, the reason for overestimating priming effect sizes is the model's abstraction from many additional relevant representations. In this case, actual selves have a much richer conceptual structure than just the single node we used in the simulation. Again, however, the important point is that the simulation generated an overall data pattern that is fairly similar to the results of Smeesters et al.'s experiments, as the following analysis shows.

We computed an analysis of variance to test the activation pattern suggested by our simulation. The results are displayed in Table 2. The pattern of significant effects matches the analyses reported by Smeesters et al. (2009) closely, as the last two columns

Table 1
Results of Simulation 5: Average Activation of the Action Node "Compete" as a Function of Business Versus Religion Prime, High or Low Social-Value Orientation (SVO), and High or Low Accessibility of Self-Knowledge (cf. Smeesters et al., 2009)

Priming condition	Accessibilty of self-knowledge			
	High		Low	
	High SVO	Low SVO	High SVO	Low SVO
Business	.64	.66	.65	.65
Religion	−.36	−.24	−.35	−.33

Table 2
ANOVA for the Activation Pattern From Table 1 (Simulation 5): Comparison With Significant Effects From Smeesters et al.'s (2009) Experiment

Source	Significance of effects in simulation		Significance of effects in original	
	df	*F*	Experiment 1	Experiment 2
Prime	1	11,856.0***	yes	yes
SVO	1	19.4***	yes	yes
Accessibility	1	5.9*	?[a]	?[a]
Prime × SVO	1	11.0***	no	yes
Prime × Accessibility	1	3.7	yes	yes
SVO × Accessibility	1	9.3**	yes	no
Prime × SVO × Accessibility	1	3.7	no	no
Residual	7,992			

Note. ANOVA = analysis of variance; SVO = social-value orientation.
[a] The original article does not report if this effect was significant.
* $p < .05$. ** $p < .01$. *** $p < .001$.

of Table 2 indicate. Main effects of priming and SVO were significant in our simulation and in both experiments from the original study. The Priming × SVO interaction, significant in our simulation, was significant in Smeesters et al.'s Experiment 2 but not in their Experiment 1. Our simulation fails to predict the two-way interaction between priming and self-concept accessibility that was found in both original experiments, but it does predict the self-concept Accessibility × SVO interaction found in the original Experiment 1. No three-way interactions were found in our simulation nor in either of Smeesters et al.'s experiments. While the correspondence between simulation and experimental results was not perfect (but neither was the correspondence between the two original experiments), the general finding was reproduced that effects of priming on behavior are constrained by self-concepts, when these self-concepts are cognitively accessible. Our model thus makes the same predictions as Wheeler et al.'s (2007) active-self account of behavioral priming effects.

Discussion of limitations. The outcome of simulations with localist constraint networks often depends on decisions related to the topology of the network. In the present model, the main constraints—the connection weights representing affective meanings of the relevant concepts—stem from independent empirical rating studies, substantially reducing opportunities for design decisions that produce desired results. However, it is clear that considerable degrees of freedom still exist.

For example, the choice of concepts for representing independent and dependent variables is not always as obvious as in Simulation 1, where the original experimental procedures left little room for interpretation, so the choice of "rude," "polite," and "interrupt" as concepts in the model suggested itself (Bargh et al., 1996). Other cases are less straightforward. For example, we chose the verbal label "Black" as input stimulus in Simulation 2, where the original experimental material had consisted of photographs of faces (Bargh et al., 1996, Experiment 3), and we chose the verbal label "compete" as output of Simulation 5, where the original dependent variable was the number of chips the original participants allocated to themselves in a game (Smeesters et al., 2009). These choices involve some degree of our own subjective interpretation of what the experimental situations might have meant to the participants. Other interpretations might be possible, and they might invoke different concepts with different affective meanings, which in turn might lead to different simulation results. However, subjective interpretation can never be disposed of entirely in explanations of social interaction, for the fundamental fact that all social interaction involves the subjective creation of meaning (Blumer, 1969). Using empirical evaluation-potency-activity profiles of concepts as input for a constraint-satisfaction algorithm provides a way of operationalizing and quantifying such meaning making, but it remains interpretative, and we may be wrong about it. At the very least, our simulations can be taken as a demonstration of plausibility: The mechanisms of parallel constraint satisfaction and affective meaning maintenance, as specified in our model, do in fact generate data patterns similar to those observed in a variety of important studies of behavioral priming.

Another example of degrees of freedom in the design of the model is the inner topological structure of the network. The connections among nodes representing the self, target person, and behavior were of course not dictated by external data, but rather depended on our decisions, which warrant some critical discussion. As explained before, the design of the model was motivated by theoretical considerations. The self-target-behavior structure was intended to reflect previous theoretical accounts of priming such as active-self, relationship-oriented, and direct-expression approaches (for review, see above or Wheeler & DeMarree, 2009). It is also compatible with Heise's (2007) affect control theory, which provided the idea of taking evaluation-potency-activity as a parsimonious, yet powerful scheme of representation that allows for direct comparison of qualitatively distinct concepts both within and across cultures. Despite these theoretical considerations, it is possible that more parsimonious variants of the model might produce similar predictions for behavioral priming. Accordingly, we repeated Simulations 1–5 with different topological versions of the model, allowing for a number of observations.

First, a direct link between the primed concept and the behavior evaluation-potency-activity patterns, included to match the model with direct-expression accounts of priming (e.g., Bargh et al., 1996), is not necessary to generate the data patterns of the simulated experiments. A model that only allows the prime to influence self and target representations performs similarly, with somewhat reduced main effects of priming in Simulations 1, 4, and 5, and

relatively more pronounced interaction effects between self-concept and priming in Simulation 5.

Second, it is possible to generate the mere pattern of main effects in Simulations 1–4 (but not the accurate base rate predictions from Simulations 1–3) with a much simpler model that does not include self and target representations. In the priming condition of the respective experiments, the concepts are simply closer to the dependent variable in EPA space, thus rendering convergence in the constraint-satisfaction algorithm easier. However, such a simpler model has obvious limitations when it comes to the more complex interaction effects between different kinds of representations, examined in Simulation 5, and our preference is for *one* model to be able to predict behavior in both simpler and more complex situations.

Third, it should be noted that we were unable to generate all of the described simulation outcomes with any model that represented less than all three of the Osgood dimensions. This is an important point, since many influential models of social perception rely on only two dimensions (e.g., Carson, 1969; Fiske, Cuddy, & Glick, 2007; Sadler et al., 2011). It was thus a theoretically plausible expectation that the model might perform equally well after cutting off connections relating to one dimension (probably either potency or activity), but this was not the case.

While all the limitations of the model discussed so far relate to decisions about its overall structure, there are further limitations when it comes to certain phenomena that are important for understanding priming, but outside the scope of our model. As already mentioned in the above discussion of Simulation 4, our model does not take into account constraints on behavior that arise from representations other than interpersonal ones. In Simulation 4, this shortcoming resulted in an overestimation of the baseline willingness of the participants to help a stranger, because the model has no representation of the cost possibly incurred by the specific kind of help in Fitzsimons and Bargh's (2003) experiment. Likewise, our model does not capture ecological constraints on automatic social behavior, which have been demonstrated recently by Cesario, Plaks, Hagiwara, Navarrete, and Higgins (2010). In the study, participants' displaying fight or flight action tendencies after exposure to a threatening prime depended on whether they were seated in an enclosed booth (where no distancing behavior was physically possible) or in an open field.

However, we hold that the failure to account for such phenomena is a result of the parsimonious structure of the specific constraint network we used, and not a problem with the underlying theoretical ideas. In principle, a parallel-constraint-satisfaction account could very well handle additional, more physical, representations. Also, the work of Smith-Lovin (1987a) pointed to many physical settings having culturally shared affective meanings comparable to the identity, trait, and behavior concepts in our simulations: just consider how you feel inclined to behave differently at a *cemetery* as opposed to a *dance club*.

Another serious limitation of our model is that it does not capture theorizing about the distinction between explicit and implicit social judgment (e.g., Cunningham & Zelazo, 2007; Fazio & Towles-Schwenn, 1999; Gawronski & Bodenhausen, 2007). Explicit evaluations of social objects reflect judgments that people consciously endorse. They are usually assessed with questionnaire items or semantic differential techniques built upon Osgood et al.'s (1957) work. Implicit judgments, in contrast, rely on associations and automatic pattern recognition. A popular method for their measurement is the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998). To be sure, neural network models that can simulate implicit as well as explicit attitudes by varying the degree of activation have been proposed (Monroe & Read, 2008; Van Overwalle & Siebler, 2005). However, in our model, the problem relates more to the input data than to the handling of activation in the constraint-satisfaction process: The evaluation-potency-activity ratings we used were compiled with the semantic differential, a technique often understood as measuring explicit attitudes, to predict automatic behaviors, assumed and shown to follow more implicit judgments (e.g., Cesario et al., 2006).

This apparent inconsistency is much less important than it might first seem, since the evaluation-potency-activity data we used in the simulations reflect a cultural consensus, rather than the sentiments of individuals (Heise, 2010). Misreporting affective reactions toward specific stigmatized groups might be a strategy employed by some participants in cultural surveys, in order not to violate their consciously endorsed values (cf. Gawronski & Bodenhausen, 2007). However, we can expect that such individual deliberate deviations from the "true" feelings are captured by the standard deviations (which we used as parameter in our simulations) but do not distort the mean estimation of culturally shared affective meaning very much (cf. Heise, 2010). This reasoning parallels Arkes and Tetlock's (2004) argument that what implicit measures of social judgment gauge is precisely shared cultural knowledge, as opposed to personal prejudice. Accordingly, we were able to precisely reproduce the results of experiments that involved priming concepts of stigmatized groups (Simulations 2 and 3) with our methodology.

Even so, some limitations still remain: It should be noted that we were unable to simulate results by Cesario et al. (2006, Experiment 2), who showed that individuals' implicit (but not explicit) attitudes toward the elderly moderate the classic effect of priming an elderly stereotype on walking speed (Bargh et al., 1996, Experiment 2).[6] There are many possible explanations for this finding, but they are hard to assess as we do not have any precise data comparable to the implicit measures used by Cesario et al. (most important, no individualized data). In an ideal world, we would have been able to use sentiment repositories compiled with implicit association test or similar techniques as input for our model. However, we are not aware of any such data, and as the overall reported simulation results show, the semantic-differential data we used produced results that are accurate enough across a wide variety of important priming experiments.

Our model also has limitations when it comes to effects of priming on more idiosyncratic representations of other people or relationship goals attached to them. On a very general level, such goals are part of the cultural stock of identity meanings, as has been demonstrated for the inclination to help friends in Simulation 4 (cf. Fitzsimons & Bargh, 2003). But friends, as well as other types of relationship partners, come in all sizes and shapes, and, depending on our individual experiences with one particular person, we may hold relationship goals far more specific than the ones implied by the general cultural stereotype of that relationship. With

[6] Our model does, however, nicely reproduce the main effects of these experiments.

our approach, we cannot model priming effects originating from such more personal representations (e.g., Fitzsimons & Bargh, 2003, Studies 4a and 4b; Shah, 2003). However, we conjecture that the information-processing mechanisms in these instances are no different from those we propose here. All that is probably needed is to replace the general cultural affective meanings we used as input in our model with more idiosyncratic representations. Parallel constraint satisfaction might be the very mechanism to explain how such individualized meanings arise from combining more general cultural meanings. The personal affective representation of a specific close friend could be an "amalgamation of affective meaning" (Averett & Heise, 1987) of all the concepts one would use to describe this particular friend (not only "friend," but also a variety of traits and past behaviors). Then, the prime-to-behavior pathway would still work in the way implied by the model in Figure 7, only that the target person representation would consist of multiple different rather than just one "proxy" node. We thus think that the failure of our model to account for such more individualized instances of priming effects is not a problem with any of the proposed mechanisms but, rather, a result of the lack of accurate data at the individual level.

Conclusion. Despite limitations, natural for any computational model, we demonstrated how the combination of affective meanings with parallel constraint satisfaction provides a compelling explanation of automatic social behavior. We used a single connectionist network, based on general principles of representation, to simulate a variety of behavioral priming experiments representing different lines of theorizing. The only changes we made in the model parameters in different simulations were given by independent empirical data from a research program devoted to assessing cultural norms of sentiments (Heise, 2010). With this strategy, we were able to reproduce the data patterns reported by the authors of the original studies. We now turn to our second model as a proof-of-principle that the proposed mechanisms can be performed by real neurons in the brain, providing the missing link between purely symbolic representations and ultimately carrying out an action in a physical world with a physical body.

Model 2: Biologically Realistic Spiking Neurons

Our second model is aimed at specifying a biologically plausible implementation of our suggested representational principles in the brain. In other words, the model to be described in this section presents a hypothesis about the neural computations underlying behavioral priming phenomena. The previous, connectionist model operated at a level that Eliasmith (in press) called shallow semantics, elucidating the mechanisms by which symbolic representations of self, targets, and primes govern the activation of symbolic action representations, mediated by affective processes represented as symbols. In contrast, here we target the deep semantics of priming: How do such, initially symbolic representations translate into patterns of activity in areas of the brain that control the motor system?

The Neural Engineering Framework. Our model is based on the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003), a set of mathematical principles developed to describe how representations and transformations of representations can be achieved by interconnected populations of spiking neurons. The NEF has been used previously to model a whole variety of cognitive tasks in a biologically plausible way. To name only a few examples, Stewart and Eliasmith (2011) proposed an NEF model of action control in the basal ganglia that successfully solves the famous Tower-of-Hanoi task (cf. H. A. Simon, 1975). Thagard and Stewart (2011) modeled creative problem solving as emergent binding in spiking neural populations. DeWolf and Eliasmith (2011) provided a detailed account of the motor system as a control hierarchy, where neural firing patterns at higher levels correspond to the more abstract representations of movements that can be decompressed into more lower level representations of single components of such movements. In the present priming model, we deal only with high-level representations of actions (such as "interrupting someone"), but the model by DeWolf and Eliasmith (2011) can be taken as proof-of-principle that the corresponding spiking patterns in motor cortex can set off and control the multiple components that are part of such an abstract action (e.g., moving the legs to approach the other person, assuming a dominant posture, raising one's voice), in accordance with the decompression mechanisms of semantic pointers (see the introduction).

The mathematical principles of the Neural Engineering Framwork have been implemented in NENGO, a software package[7] that can be used to create spiking-neuron models in a drag-and-drop manner. NENGO comes with a library of standard components that automatically set up biologically realistic simulations of spiking neurons. Parameters can be adjusted to allow for different types of neurons and neurotransmitters and thus maximize the biological plausibility of a model. Our present, highly simplified model is intended only as a proof-of-principle, not a detailed proposal concerning all the brain activity underlying the automaticity of social behavior. Therefore, we simply used the standard neuron model of NENGO, the Leaky Integrate-and-Fire neuron (LIF), leaving all the default parameters. The model works as follows. Current flowing into a LIF neuron affects the voltage. Upon reaching a voltage threshold, the neuron fires and thereby induces a flow of current to all the connected neurons. All the parameters of the LIF neuron model have been set in order to be consistent with neurophysiological data (for details, see Eliasmith & Anderson, 2003).

Concepts as patterns of neural activity. In the localist model above, abstract symbolic concepts such as traits or stereotypes were represented as single nodes in the network. How can they be represented in a spiking neuron model created with NENGO? An example is displayed in Figure 9, where the concepts "rude" and "polite," i.e., the priming conditions in Bargh et al.'s (1996) Experiment 1 are represented as unique patterns of spiking activity in a population of 300 simulated neurons. This view of concepts represented in the brain as distributed patterns of neural activity is in line with our recent suggestions about the nature of scientific concepts, intentions, and emotions as semantic pointers (Blouw et al., 2012; Schröder, Stewart, & Thagard, 2012; Thagard, 2012a; Thagard & Schröder, 2012). It also fits with technological advances in neuroscience related to the use of pattern analysis to infer specific thoughts or mental states from fMRI or EEG recordings (e.g., Haynes et al., 2007; Shinkareva et al., 2008). In the

[7] The software package, description, tutorials, and a database of previous models can be downloaded from http://www.nengo.ca

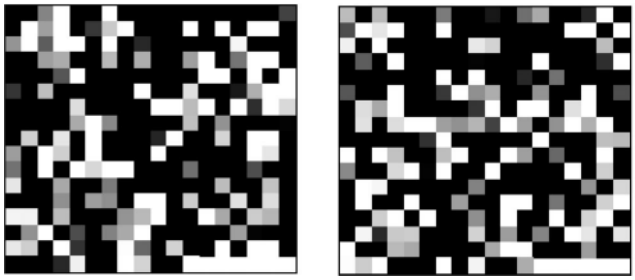

Figure 9. Representation of symbolic concepts as patterns of spiking activity in neural populations. The small squares represent individual neurons; their brightness corresponds to their voltage at a given point in time. White squares are neurons currently spiking. In this example, the activity pattern of the left population represents the concept "rudeness"; the pattern to the right "politeness." These models were generated with the NENGO simulation program.

NENGO simulation environment, neural populations can be assigned a whole vocabulary by defining specific spiking patterns that map onto specific concepts. This assignment can either be implemented through a random procedure, i.e., random patterns are generated and associated with specific concepts, or through having the relations between different activity patterns capture the semantic relations between the associated concepts.

In our present model, the semantic relations among concepts are given through their similarity along the evaluation-potency-activity dimensions, in line with the affective-meaning-maintenance mechanism reviewed above and shown to account for behavioral priming effects with our first, connectionist model. Based on neurophysiological evidence from studying motor and visual cortex (e.g., Georgopoulos, Schwartz, & Kettner, 1986), artificial neurons in NENGO have preferred direction vectors, i.e., they fire more strongly the more similar the orientation of a stimulus in the represented space is to the orientation to which the neuron is most sensitive. In our model, the space is not to be understood literally as in the case of visual representation of the environment, but rather in terms of the basic affective dimensions of Osgood et al. (1975). The spiking behavior of each neuron, represented in Figure 9 by the brightness of the squares, is contingent upon the closeness of the represented concept to the "preferred" evaluation-potency-activity (EPA) profile of the neuron (for mathematical details, see Eliasmith & Anderson, 2003). Therefore, concepts that are semantically similar in terms of their placement in the affective space will produce similar overall patterns of distributed spiking activity in an artificial neural population generated with NENGO. As in the previous, connectionist model, EPA profiles for concepts are given by the existing empirical data from cultural surveys (cf. Heise, 2010). The activation of semantically similar concepts through the spread of activation mediated by central EPA nodes in the localist model thus corresponds to the elicitation of similar spike patterns in the neurocomputational model, mediated by a fixed sensitivity of individual neurons to certain EPA configurations.

Model architecture. Conceptually, the components of this model are largely the same as in the localist model described in the previous section. The input is given by activated representations of the self, a target person of the social interaction, and the prime. The output is a representation of social action corresponding to the dependent variable of the simulated experiment. However, we modeled these representations as activity patterns in neural populations of 300 neurons each. We also tentatively mapped these neural populations onto different anatomical brain areas thought to be involved in the processing of the relevant information. Naturally, this anatomical mapping is highly simplified and by no means do we wish to make any strong claims regarding the brain areas possibly involved in the control of automatic social behavior. The existing knowledge is insufficient to credibly engage in such a complex endeavour, given how young the field of social neuroscience is (for a state-of-the-art overview, see Todorov, Fiske, & Prentice, 2011). A complete, neuroanatomically correct model would have to integrate findings from areas such as social judgment (e.g., Cunningham & Zelazo, 2007), emotion processing (e.g., Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012), and motor abstraction (e.g., Gallese, 2009), which are hardly understood on their own. What we intend here is to show that the mechanisms and computations we propose to be crucial to the control of automatic social behavior can be performed by more realistic neurons. Since the representations in our second model are highly distributed, they need not be confined to single brain areas but could in principle be extended over multiple networks. For parsimony, we currently restrict the model to the architecture displayed in Figure 10.

In our model, representations of self and target persons are given by activity in the medial prefrontal cortex (Amodio & Frith, 2006; Denny, Kober, Wager, & Ochsner, 2012; Mitchell, Macrae, & Banaji, 2006; Van Overwalle, 2009). In all the priming experiments reviewed above, the procedure involves visual presentation of the prime; it is thus reasonable to assume that the prime is first represented by a spike pattern in visual cortex, displayed on the left side of Figure 6. We assume that this activity immediately projects to the amygdala (and, most probably, other brain areas involved in the processing of affect

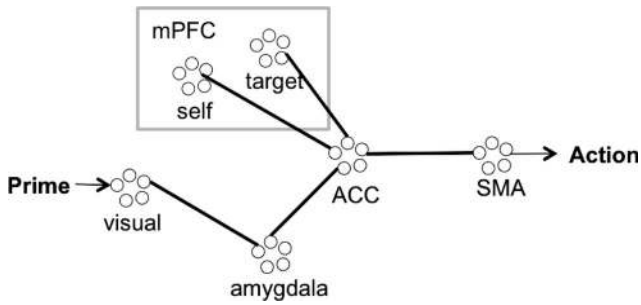

Figure 10. Spiking-neuron model of behavioral priming effects. Representations of self and interaction target person are semantic pointers (= patterns of activity) in populations of neurons in medial prefrontal cortex (mPFC). Perception of a primed concept in visual cortex causes activation in the amygdala (and other evaluation-related brain areas such as insula or the dopamine system, not displayed here for parsimony). Anterior cingulate cortex (ACC) is a candidate structure for computing the binding of activity patterns that represent affective meanings of self, target, and prime. The resulting activity gets projected to a brain area representing motor vocabulary (possibly, the supplemental motor area [SMA]) and activates the most semantically related action representation.

such as insula or the dopamine system; for reviews, see Cunningham & Zelazo, 2007; Lindquist et al., 2012; Thagard & Aubie, 2008). This direct link in our model is consistent with EEG data showing very quick responses of the brain to emotion-laden verbal stimuli (e.g., Hofmann, Kuchinke, Tamm, Võ, & Jacobs, 2009; Skrandies, 1998). We also need a neural population that computes the integration of the single representations of primed concept, self, and target person, similar to the parallel constraint satisfaction in our localist model. Based on Tsakiris and Haggard's (2010) review of the neural structures underlying the control of action, we think that anterior cingulate cortex (ACC) is plausibly involved in performing that task (see also Cunningham & Zelazo, 2007). Finally, we suggest that the resulting holistic pattern of activity activates a semantically related abstract representation of action in the motor areas of the brain (possibly, the supplemental motor area [SMA], cf. Tsakiris & Haggard, 2010). This idea of abstract motor representation fits with claims about the existence of a nonverbal action vocabulary in the motor system, consisting of high-level neural models of underlying more specific motor programs in relation to goals (Arbib, 2005; Fogassi, 2011; Gallese, 2009). As mentioned earlier, the hierarchical model of motor control by DeWolf and Eliasmith (2011) specified how such high-level motor abstractions unfold into their single components and, ultimately, in movements of the muscles. That level of detail is far beyond the scope of our present social psychological model of priming effects, but the connection is feasible in principle, since both the low-level account of motor control by DeWolf and Eliasmith (2011) and our present high-level account of automatic social action are based on the same neural mechanisms and operationalized within the same modeling environment.

Simulation 6. In this section, we describe our simulation of the classic demonstration of behavioral priming effects by Bargh et al. (1996), comparable to our Simulation 1 with the localist constraint network described above. We intend to demonstrate that the mechanisms we proposed as guiding automatic social behavior can also be performed by the much more biologically detailed (yet, also much more complicated) NENGO model, thus enhancing the neuroscientific credibility of our account.

In Bargh et al.'s (1996) experiment, subjects primed with rudeness were more likely to interrupt the experimenter from speaking to a third person than subjects primed with politeness. As in the connectionist model, input and output of our present simulation were based on empirical evaluation-potency-activity ratings of "Myself as I really am" (self), "student" (target), and "rude" versus "polite" (experimental priming conditions) from Francis and Heise's (2006) repository of cultural sentiments for the United States. For technical reasons, these were replicated to nine dimensions and then transformed into unit vectors. These vectors were used by NENGO to simulate semantically meaningful spiking patterns in the respective neural populations displayed in Figure 10 (see Figure 9 again for a visualization of the different spike patterns corresponding to the primes). To this end, NENGO randomly created preferred direction vectors for each of the 300 neurons in every neural population, causing the neurons to fire more often the closer the vector representing cultural sentiments was to their preferred direction vector. Since NENGO keeps the preferred direction vectors fixed regardless of which concept is currently represented, any neural population will exhibit similar overall firing patterns whenever similar concepts are represented (for mathematical details, see Eliasmith & Anderson, 2003).

In the neural population representing a circuit from the supplemental motor area (termed SMA in Figure 10), we implemented a "one-item action vocabulary," operationalized as the spike pattern corresponding to the vector we created out of the empirical evaluation-potency-activity profile of the verb "interrupt" from the Francis and Heise (2006) data set. The critical question for the simulation of Bargh et al.'s (1996) experiment is as follows: When the incoming activity from ACC causes the neurons in SMA to fire, how similar is the resulting firing pattern to the pattern representing "interrupt"? In order to match the original experimental result, this similarity should be much higher when the spike pattern corresponding to "rude," as opposed to "polite," is provided as input.

Assessing the semantic similarity of NENGO spike patterns involves two steps. First, the spiking activity of the neural population has to be decoded into the corresponding numerical vector, using the mathematical methods developed by Eliasmith and Anderson (2003). Second, that vector must be compared with the target vector (in this case, a transformation of the evaluation-potency-activity profile of "interrupt"). A metric to assess the similarity of two vectors is the dot product (since the vectors are standardized, this is comparable to the Pearson correlation coefficient in statistics). In our simulation, the dot product fluctuated between .60 and .62 when "rude" was the prime, and between –.33 and –.29 when "polite" was the prime (overtime fluctuations are due to noise in the simulated neurons). The spike pattern generated in the SMA population was thus relatively similar to the pattern corresponding to "interrupt" in the rudeness priming condition, and quite dissimilar in the politeness condition. This result matches Bargh et al.'s (1996) finding that the subjects were more likely to interrupt the experimenter after exposure to the rude-related stimuli, and less likely to do so when primed with concepts related to politeness.

Discussion. Our model and simulation support our claim that the principles we have proposed to explain the effects of priming on subsequent social behaviors are compatible with existing knowledge about neural mechanisms that underlie both the representation of symbolic concepts and the control of movement. According to the semantic pointer hypothesis (Eliasmith, in press), the brain represents concepts as patterns of activity that are meaningfully constrained by related underlying representations. Most important, semantic pointers are multimodal and can thus direct flows of information from verbal representations, like primed concepts, to the affective reactions and motor programs influenced by those.

In our simulation, we demonstrated how the semantic properties of concepts implied by their location in Osgood's affective space, indicated through the ratings of cultural informants in empirical surveys, can be encoded in the spiking behavior of neurons. The parallel satisfaction of multiple constraints resulting from various representations active at the same time, the core mechanism of our first, localist model of priming effects described above, can also be computed by spiking neurons as was also demonstrated by Thag-

ard and Aubie (2008).[8] The resulting, holistic pattern of activity still carries semantic information which, in our simulation was sufficient to generate an action prediction corresponding closely to the one resulting from the simpler, localist model and the experimental data from Bargh et al. (1996). Table 3 summarizes the correspondence between the components and mechanisms of the localist and the neurocomputational model.

Regarding limitations of the model, we already pointed to the high degree of simplification of the anatomical mappings provided in Figure 10. Patterns of neural activity representing the self, another person, a prime, and the affective meaning of these entities are certainly not confined to a single, easily localizable region of the brain. However, providing a complete anatomically realistic account of how the brain controls automatic social behavior was not our goal. Given the highly distributed nature of representations in general and representations of affect in particular (see Lindquist et al., 2012), such an endeavor would require inclusion of almost the whole brain in the model. This would make it both hard to understand and probably wrong, given the limited existing knowledge about the social brain. Rather, our goal was to demonstrate how compatible our mechanistic account of priming is with state-of-the-art theorizing about the computational processes going on in realistic neurons. We specified how spiking neurons can represent affective meaning and how they might integrate different representations to generate semantically related actions in the motor system.

General Discussion

Our present work was motivated by the goal to provide a unified explanation for empirical data showing that the mere activation of a concept in the mind of a person can cause them to align their subsequent behavior automatically with the meaning of that concept (for reviews, see Bargh, 2006; Dijksterhuis & Bargh, 2001; Wheeler & DeMarree, 2009). To this end, we proposed three general principles of representation and processing (parallel constraint satisfaction, affective meaning maintenance, and semantic pointers) that we think can account for the abundant findings in the literature on behavioral priming. To support our claim, we developed a localist network model implementing the first two principles, and we showed in a series of simulations that it reproduces major experimental results. In a complementary neurocomputational model, we demonstrated how the mechanisms can be implemented by populations of spiking neurons, thereby enhancing the neurological plausibility of our account. In the remainder of this article, we first discuss the potential of our theory to integrate previous, sometimes competing approaches to explaining priming-to-behavior effects. Then we defend our view that full explanations of social behavior require theoretical integration across the psychological, cultural, and neural levels of explanation, as in our present attempt to elucidate priming. Finally, we discuss some open questions and point to further research needs.

How Our Model Integrates Previous Explanations for Priming Effects

Theoretical explanations of automatic social behavior, reviewed in the introduction of this article, include accounts of a perception-behavior link (e.g., Bargh et al., 1996), interaction goals (e.g., Cesario et al., 2006; Fitzsimons & Bargh, 2003), the active self (Wheeler et al., 2007), situational misattribution (Loersch & Payne, 2011), and complex metaphorical structures (Bargh, 2006). We think that our view is compatible with all of these ideas and has the potential not only to integrate them but also to make them more rigorous through a computational formalization of the proposed mechanisms. The latter aspect is especially important since priming researchers have often felt unable to predict which of the many possible effects of a single priming procedure will actually occur in a given situation (the one-prime-many-effects problem, see Bargh, 2006), whereas the computational models we have proposed make precise predictions.

According to Bargh et al.'s (1996) original introduction of behavioral priming effects, they are caused by acquired associations between situational features and behavioral responses. The activation of stereotypes and traits through a priming procedure renders semantically related behavioral representations more accessible. The accessibility of these behavioral tendencies makes the actual behaviors more likely to occur, without any conscious decision required. The latter reasoning fits with the famous principle of ideo-motor action suggested by William James (1950/1890) and modern neuroscientific views of symbolic representations as rooted in lower level sensorimotor representations (e.g., Barsalou, 1999). Eliasmith (in press) described the relationship of symbols and related sensorimotor experience as the interplay of shallow vs. deep semantics. He also developed the semantic pointer hypothesis of cognition, a mathematically formalized account of how populations of spiking neurons can perform the necessary computations that underlie the activation of behavioral representations following the activation of semantically related symbolic concepts. We demonstrated how these semantic pointer principles can be used for an accurate, biologically plausible computer simulation of the groundbreaking experiment by Bargh et al. We also precisely specified the semantic relations between trait, stereotype, and behavior representations by mathematically treating them as locations in the affective evaluation-potency-activity space (Osgood et al., 1975), based on independent data from cultural surveys (cf. Heise, 2010).

A somewhat alternative explanation of priming effects emphasizes the role of motivation, as opposed to mere concept activation (Cesario et al., 2006). According to this view, priming a stereotype activates the interaction goals a primed person has toward members of the stereotyped category. Similarly, social roles and types of relationship entail specific interaction goals (e.g., Fitzsimons & Bargh, 2003; Shah, 2003). Our own theory of automatic social behavior is compatible with this interaction goal account. We think

[8] In our model, the neural population representing ACC simply summed up the inputs from the prime, self, and target spike patterns. If we understand the integration of patterns as neural binding, our present operationalization is reminiscent of a simple synchronization account of binding (Von der Malsburg, 1981). For many higher level cognitive operations, this is inadequate (Jackendoff, 2002); therefore, Eliasmith (in press) generally used the much more sophisticated function of circular convolution to simulate binding in spiking neurons (cf. Plate, 2003). Yet we chose to implement the most simple mechanism possible in our model in the interest of parsimony. Besides, there is no reason why some processes in the brain should not operate on a very simple associative basis, while others might still require the implementation of more complex mathematical functions in the connection weights between different neural populations.

Table 3
Correspondence of the Localist and Neurocomputational (NENGO) Models of Priming

Element/process	Localist model	NENGO model
Symbolic representation of prime, self, target, and behavior	Single nodes	Spike patterns in neural populations
Representation of affective meaning (EPA)	Connections with special EPA nodes	Preferred direction vectors of neurons
Activation of semantically similar concepts	Activation spreads via direct connection	Generation of similar spike patterns
Parallel constraint satisfaction	Interactive competition in whole network	Pattern aggregation in ACC population
Action prediction	Activation of single, symbolic action node	Similarity of SMA spike pattern to action semantic pointer

Note. EPA = evaluation-potency-activity; ACC = anterior cingulate cortex; SMA = supplemental motor area.

that such goals are not static representations requiring nodes on their own in the localist network displayed in Figure 2 but are rather an emergent property of the dynamic parallel-constraint-satisfaction process. In our model, a relationship is given by a configuration of self and target representations, each of which carries a specific affective meaning inherent in the verbal concepts people would use to describe this self–other relation. The effect of priming is that it renders a concept more capable of influencing those interpretations of the relationship. Parallel constraint satisfaction determines the most likely action to follow from such dyadic representations, as if there were something like the activation of an explicit relationship goal. For example, in our Simulation 4, we showed that the model needs no explicit representation of a goal to help friends in order to produce the output that this is a very likely action. Rather, this information is contained in subtle form in the culturally shared affective meaning of a "friend" in relation to oneself, and the action emerges automatically from the computation of constraint satisfaction. There is, thus, no need to assume some kind of dichotomy (implied by Cesario et al., 2006) between "perception-behavior-link" and "motivational" accounts of priming. Rather, the same mechanism—satisfying the constraints given by multiple affective meanings—comes in different guises, more direct-expressive or more motivational, depending on the situation.

A further important line of research under the behavioral priming paradigm is related to the active-self account (Wheeler et al., 2007), according to which the activation of self-knowledge is a key mechanism causing the link from prime to action. Both of our models are coherent with this view. This fit becomes apparent from the self-representation we included in all of our models, sometimes in the default "Myself as I really am" operationalization, sometimes with a more specific label, as implied by the respective original studies. We think that our contribution to the active-self paradigm lies in the formalization of the mechanism by which the self-concept moderates individuals' susceptibility to priming. In our Simulation 5, the specification of self-meanings and primed concepts with Osgood's measurement scheme and the parallel-constraint-satisfaction algorithm led to a fairly close replication of the complex data patterns reported by Smeesters et al. (2009). Parallel processing naturally deals with situational knowledge integration proposed to be key for priming by Wheeler et al. (2007).

Loersch and Payne (2011) proposed that all effects of priming result from misattributing the content of a prime to the objects of the environment that are currently in the focus of the primed person's attention. Our theory has substantial similarity with the misattribution model. Like Loersch and Payne, we assume that the effects of priming on behavior are not immediate, but stem from a higher accessibility of the primed concept, which is then likely to bias people's representation of the situation. We propose that what Loersch and Payne called misattribution is the result of parallel constraint satisfaction. In some ways, our approach is narrower than the misattribution approach, since we target only behavioral priming, whereas Loersch and Payne proposed that all forms of priming result from the same mechanism. However, we already made clear that we believe that parallel constraint satisfaction can probably explain all kinds of priming (see the introductory section on mechanisms). Moreover, numerous studies showed that our basic principles of parallel constraint satisfaction and affective meaning maintenance both provide compelling explanations of impression formation and attribution (e.g., Averett & Heise, 1987; Kunda & Thagard, 1996; Read & Miller, 1993; Schröder, 2011; Smith-Lovin, 1987b), another class of phenomena subject to priming phenomena (see Loersch & Payne, 2011, for review). It is thus likely that the explanatory power of our present models extends to a wider range of priming phenomena, although we have chosen to focus on behavioral priming here. To summarize, we believe that our present work takes the unifying approach by Loersch and Payne further by spelling out the crucial information-integration mechanism with the rigorousness innate to a computational model.

Finally, our approach substantiates theoretical reflections by John Bargh (2006) on "what [we have] been priming all these years" (p. 147). He suggested turning attention to language, in order to understand how complex conceptual structures, as opposed to single concept–behavior links, govern the operation of priming effects. According to this view, priming procedures influence the metaphorical perspectives people assume in order to make sense of the world (cf. Lakoff & Johnson, 2003). Drawing on the extensive work by Heise and colleagues (e.g., Heise, 1979, 2007, 2010), we showed that the technique of measuring affective meaning developed by Osgood et al. (1957, 1975) offers the possibility of a precise mathematical operationalization of the conceptual structures that guide the control of social interaction. In fact, a preoccupation with metaphor was precisely the starting point of Osgood's endeavours (cf. Miron, 1969; Osgood et al., 1957). Verbal concepts denoting stereotypes, traits, behaviors, etc. are meaningfully distributed over the affective space. As we have demonstrated with our first, connectionist model, the data patterns of priming experiments follow those affective arrangements of concepts. In our model, key results, such as interrupting someone

following a rude prime, helping someone after a friend prime, and keeping more chips for oneself in a dictator game following a manager prime, all resulted from the evaluation-potency-activity configurations of these concepts. As we have further shown with our second, neurocomputational model, such a conceptual structure in affective space can be plausibly represented by spiking neurons in brains. We thus fully agree with Bargh (2006) that behavioral priming effects follow the complex semantic structures objectified in language (Berger & Luckmann, 1966), and we have proposed precise mechanisms for how these structures translate into action in specific situations.

Multilevel Mechanisms in the Explanation of Social Behavior

We believe that a theoretically rich understanding of social behavior requires attention to multiple levels of explanation that have been the focus of different scientific disciplines. Sociologists and anthropologists have emphasized the role of socially constructed patterns of meaning that structure social interaction. Psychologists have looked at cognitive and affective processes that underlie the regulation of behavior. Finally, neuroscientists have started to understand the mechanisms of information processing in the brain that perform the computations necessary for a biological system to control social action. Our theory of automatic social behavior draws on insights from all these disciplines.

From sociology and anthropology, we took the idea of culture as stable cognitive–affective structures shared across the minds of culture members (DiMaggio, 1997; Heise, 2010; Romney, Boyd, Moore, Batchelder, & Brazill, 1996). We used mean empirical ratings of relevant concepts along the evaluation-potency-activity dimensions from existing cultural surveys to determine the inputs and outputs of our simulations. Heise (2010) empirically analyzed the tremendous intracultural reliability and overtime stability of such data. Furthermore, it was shown in various studies that cross-cultural variations in these ratings can be tied to existing knowledge about the characteristics of the cultures involved (e.g., Schneider, 2004; Schröder, Rogers, et al., 2012; H. W. Smith, Matsuno, & Ike, 2001). Hence, evaluation-potency-activity ratings reflect a cultural consensus about the semantics of social interaction. Our ability to reproduce the results of various experiments with our model is consistent with this assumption. The participants in these experiments and the respondents in the cultural surveys we used as database for our model present very different samples, and many years lay between the respective data collections. Yet, having grown up in the same culture and sharing the same language,[9] they agree, without ever having met as individuals, about the semantic relations among stereotypes, traits, and behaviors. These relations drove the evaluation-potency-activity ratings of concepts in one sample and the automatic behaviors following priming with these concepts in the other sample.

The idea that the interaction order of societies is crystallized, maintained, and transformed in the relations of linguistic symbols has long been prominent in the influential symbolic interactionist tradition of sociology (e.g., Berger & Luckmann, 1966; Blumer, 1969; MacKinnon, 1994; Mead, 1934). What our present research suggests is that the social order is not only at work in the more deliberate actions under conscious control but also in the more subtle, automatic behaviors as they have been described in the work of John Bargh and others.

With regard to psychology, we built upon a vast body of research that can be summarized in the form of two different conceptions of how the mind works. First, our approach incorporates the principle of parallel constraint satisfaction which is a pervasive mechanism of information processing that reflects the classic ideas of Gestalts, evaluative balance, and cognitive consistency (e.g., Heider, 1946; Festinger, 1957; Rumelhart, McClelland, & the PDP Research Group, 1986; Thagard, 2000; for reviews, see Read & Simon, 2012; D. Simon & Holyoak, 2002). Our present research suggests that one of the basic mechanisms underlying priming is no different from what has been described as governing all kinds of mental phenomena from basic letter and word recognition (McClelland & Rumelhart, 1981) to decision making (Thagard & Millgram, 1995), attitudes (Monroe & Read, 2008), and person perception (Kunda & Thagard, 1996; Read & Miller, 1993).

The three-dimensional representation of social concepts and behaviors, which is key to both our models of priming can be related to a second influential psychological idea. In an extensive review of literature on both verbal and nonverbal communication, emotion, person perception, personality, and interpersonal behavior, Scholl (in press) found that positive versus negative, weak versus strong, and active versus passive dimensions, similar to Osgood's (1969) evaluation, potency and activity are ubiquitous in human experience of social relations across all cultures studied so far. He interpreted them as a "universal socio-emotional space that corresponds to an evolutionary need for coordination between individuals" (p. 1). Osgood (1969) himself has linked his dimensions to evolutionary pressures in a way that fits with contemporary appraisal theories of emotion (e.g., Oatley, 1992; Scherer, Schorr, & Johnstone, 2001), stating that it is beneficial for an organism to be able to come up with quick judgments related to the potential harmfulness, powerfulness, and activity of an aggressor (see also Scherer, Dan, & Flykt, 2006; Fiske et al., 2007).

Discussing the biological plausibility of the affective mechanisms at the core of our model of priming leads us to the neural level of explanation, or the hard-wired mechanisms of automatic social behavior. The central role we assigned to affective meaning in mediating the concept-behavior link loosely parallels more biologically oriented theories of emotion that emphasize the relatively fixed, universal behavioral responses thought to quickly follow from specific affective states (e.g., Cannon, 1929; LeDoux, 1996; Panksepp, 2000). Our highly simplified neural architecture displayed in Figure 10 implies a quick associative generation of behavior through processing a stimulus in neural networks in the amygdala, like the neural basis of the fear response studied extensively by Phelps and LeDoux (2005). Of course, the behaviors under study in priming research have complex, socially constructed meanings, so their affective significance in the conscious experience of the experimental subjects is much smaller than the

[9] Simulation 5 is an exception in that we used German-language data to simulate Dutch-speaking participants' behaviors. However, as mentioned in the description of that simulation, one can argue that Dutch and German culture are more similar to each other than to Anglo-Saxon culture (see Gupta et al., 2002). Accordingly, it is extremely unlikely that our simulation would have reproduced the complex data pattern of Smeesters et al.'s (2009) experiments by pure chance.

reactions elicited by patterns of snakes or roaring lions. Priming researchers have repeatedly demonstrated that effects are not attributable to conscious differences in mood reported by participants across priming conditions (e.g., Bargh et al., 1996; Zemack-Rugar, Bettman, & Fitzsimons, 2007). But not all affective processes are necessarily conscious, since they compete with other mental processes for the limited resources of working memory (Berridge & Winkielman, 2003; Lindquist et al., 2012; Thagard & Aubie, 2008). Our perspective, in line with Eliasmith's (in press) semantic pointer hypothesis, is that priming effects involve compressed, shallow representations of the innate, more reflex-like emotion-action pathway. Osgood's (1969) notion of affective meaning points to the ultimately biological constraints on conceptual structures and thus opens up a perspective for reconciling the view of universal brain mechanisms with the notion of culturally constructed symbolic structures as guiding the control of action.

Limitations and Directions for Future Research

The present work should be expanded in the future to address some of the limitations of our approach concerning both the development of potentially more sophisticated models as well as new experimental studies in order to test some implications of our work.

The importance of culture is a core aspect of our approach, and a natural consequence is the prediction of cross-cultural differences in behavioral priming effects. In principle, the existence of such differences has been demonstrated recently by Wheeler, Smeesters, and Kay (2011). In a social dilemma game, Chinese-born participants reacted differently to competition vs. cooperation primes than Dutch-born subjects. The lack of a Chinese sentiment repository containing the relevant concepts from Wheeler et al.'s experiment prevented us from attempting a simulation of their study. But when we developed our above-mentioned Simulation 5 of a similar experiment by Smeesters et al. (2009), we certainly encountered the cross-cultural issue. According to the German evaluation-potency-activity database we used for our simulation, concepts related to competition are rather negatively evaluated, whereas in the United States cultural repository by Francis and Heise (2006), these concepts have positive affective meanings. As we argued above, the German data probably provide a better cultural match for the Dutch-speaking participants in the original experiment (in personal communication, April 27, 2011, Smeesters agreed; see also Gupta et al., 2002). Accordingly, the simulation result matched the actual results of the experiment much better than a different, tentative simulation with U.S. data as input.[10] However, this issue should be examined much more rigorously. Our present connectionist model could be used with input data from different cultures to derive precise predictions of cross-cultural differences in the susceptibility to priming. These predictions should then be tested in experimental studies with subjects from the corresponding cultures.

If culture does constrain the automaticity of behavior through structures of affective meaning embedded in connection weights between neurons, then explanation is required concerning how brains learn these connection weights in socialization with little direct supervision. In our models, we have treated culture as simply externally given and already embedded in the neural networks. Reality is much more dynamic. Sociologists have described social interaction as a constant interplay of learning about social norms and meanings, adapting one's behavior to these meanings, and changing and renegotiating them in the course of communication (e.g., Berger & Luckmann, 1966; Blumer, 1996; MacKinnon, 1994). To our knowledge, little is known about how to link psychological and neural mechanisms such as conditioning, mere exposure, and individual cognitive appraisal to the social dynamics of socialization, creation of meaning, and cultural change. Our present models cannot handle these questions, since they are unable to learn; their connection weights are simply given. However, future extensions incorporating learning rules (e.g., see Monroe & Read, 2008; Stewart, Bekolay, & Eliasmith, 2012; Van Overwalle, 1998) might be able to elucidate how experiencing social actions influences EPA representations of concepts as much as EPA representations of concepts influence social actions. Multiagent models, simulating communication between multiple virtual agents in artificial societies (e.g., Bonabeau, 2002; Van Overwalle & Heylighen, 2006) might one day shed light on how stable, consensual structures of affective meaning are generated and maintained in cultures.

Getting back to the psychological level of explanation, another area for future research is the development of more sophisticated models that can handle temporal dynamics of priming. For example, Cesario et al. (2006, Experiment 3) demonstrated that the effects of priming of the elderly concept disappear when participants are given the opportunity to provide a short story about interacting with an elderly man. They explain this in terms of "postfulfillment inhibition" (Förster, Liberman, & Higgins, 2005): Having satisfied their interaction goal symbolically by providing the write-up, the accessibility of the elderly concept in the subjects' minds was reduced. In contrast, a simple cognitive activation account would have predicted the writing task to even further increase, not decrease, the effects of the priming procedure. Our models presented here are too simple to account for such a finding, due to the absence of temporal feedback loops. In principle, however, the Neural Engineering Framework (Eliasmith & Anderson, 2003), which we used for our neurocomputational model, provides methods for dealing with more dynamical aspects of representations (e.g., Stewart & Eliasmith, 2011). It should thus be feasible to extend our second model to handle time-sensitive processes such as the sequences of different priming methods employed by Cesario et al. (2006).

Further activities should also address the integration of more deliberate actions in the models proposed here. The impressive evidence for automaticity of behavior notwithstanding, social action is also often highly deliberate and planned (e.g., Fishbein & Ajzen, 2010). In fact, thinking about a duality in the nature of human behavior as unconscious and passion-driven versus deliberate and intentional is reflected in many contemporary

[10] The simulation of Smeesters et al.'s (2009) experiments with input data taken from Francis and Heise's (2006) U.S. data set reproduced the main effects of business versus religion priming, but not the pattern of interaction involving low versus high SVO self-concepts.

dual-process theories in social psychology (e.g., Deutsch & Strack, 2006; Fazio & Towles-Schwenn, 1999; Lieberman, 2003; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). We surmise that the basic principles we described here as an explanation of priming might as well account for an integration of more deliberate thought as an origin of behavioral control. Explicit reflection on goals and intentions is bound by the conceptual categories of language. Deliberate intentions might simply arise from binding together a multitude of concepts in working memory as a result of conscious effort (Schröder, Stewart, & Thagard, 2012). The same mechanism of parallel constraint satisfaction described here might then cause a holistic affective state to emerge from the amalgamated affective meanings of all the concepts used for generating the intention. Cunningham and Zelazo (2007) have proposed the idea of an iterative cycle of evaluative reprocessing, according to which conscious reflections about goals can override initial, automatic appraisals performed in the affective circuits of the brain. Such an overriding mechanism might be subject to the same principles of integrating different affective meanings through parallel constraint satisfaction that we have proposed here to guide purely automatic priming effects. It would be an important step toward a comprehensive explanation of social behavior in general to build a model to substantiate and test our conjecture.

To conclude, we think that our explanation of behavioral priming reflects key insights about representation and social interaction from multiple scientific disciplines. Our integration of knowledge across the cultural, social, psychological, and neural levels of explanation gives rise to a comprehensive understanding of a fascinating social psychological phenomenon.

References

Amodio, D. M., & Frith, C. B. (2006). Meeting of the minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277. doi:10.1038/nrn1884

Andres, M., Olivier, E., & Badets, A. (2008). Action, words and numbers: A motor contribution to semantic processing? *Current Directions in Psychological Science, 17,* 313–317. doi:10.1111/j.1467-8721.2008.00597.x

Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences, 28,* 105–124. doi:10.1017/S0140525X05000038

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the implicit association test?" *Psychological Inquiry, 15,* 257–278. doi:10.1207/s15327965pli1504_01

Averett, C. P., & Heise, D. R. (1987). Modified social identities: Amalgamations, attributions, and emotions. *Journal of Mathematical Sociology, 13,* 103–132. doi:10.1080/0022250X.1987.9990028

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology, 36,* 147–168. doi:10.1002/ejsp.336

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462–479. doi:10.1037/0003-066X.54.7.462

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230–244. doi:10.1037/0022-3514.71.2.230

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577–660.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London, England: Routledge.

Bechtel, W., & Abrahamson, A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks* (2nd ed.). Oxford, England: Blackwell.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge,* Garden City, NY: Anchor Books.

Berridge, K., & Winkielman, P. (2003). What is an unconscious emotion? (The case for unconscious "liking"). *Cognition & Emotion, 17,* 181–211. doi:10.1080/02699930302289

Blouw, P., Solodkin, E., Eliasmith, C., & Thagard, P. (2012). *Concepts as semantic pointers: A theory and computational model*. Unpublished manuscript, University of Waterloo.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Berkeley, CA: University of California Press.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 99,* 7280–7287. doi:10.1073/pnas.082080899

Cacioppo, J. T., Berntson, G. G., Sheridan, J. F., & McClintock, M. K. (2000). Multilevel integrative analyses of human behavior. Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin, 126,* 829–843. doi:10.1037/0033-2909.126.6.829

Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear, and rage*. Oxford, England: Appleton. doi:10. 1097/00007611-192909000-00037

Carson, R. C. (1969). *Interaction concepts of personality*. Chicago, IL: Aldine.

Cesario, J., Plaks, J. E., Hagiwara, N., Navarrete, C. D., & Higgins, E. T. (2010). The ecology of automaticity: How situational contingencies shape action semantics and social behavior. *Psychological Science, 21,* 1311–1317. doi:10.1177/0956797610378685

Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology, 90,* 893–910. doi:10.1037/0022-3514.90.6.893

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

Crawford, L. E. (2009). Conceptual metaphors of affect. *Emotion Review, 1,* 129–139. doi:10.1177/1754073908100438

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences, 11,* 97–104. doi:10.1016/j.tics.2006.12.005

Damasio, A. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation, 1,* 123–132. doi:10.1162/neco.1989.1.1.123

DeMarree, K. G., & Loersch, C. (2009). Who am I and who are you? Priming and the influence of self versus other focused attention. *Journal of Experimental Social Psychology, 45,* 440–443. doi:10.1016/j.jesp.2008.10.009

DeMarree, K. G., Wheeler, S. C., & Petty, R. E. (2005). Priming a new identity: Self-monitoring moderates the effects of nonself primes on self-judgments and behavior. *Journal of Personality and Social Psychology, 89,* 657–671. doi:10.1037/0022-3514.89.5.657

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24,* 1742–1752. doi:10.1162/jocn_a_00233

Deutsch, R., & Strack, F. (2006). Duality models in social psychology: From dual processes to interacting systems. *Psychological Inquiry, 17,* 166–172. doi:10.1207/s15327965pli1703_2

DeWolf, T., & Eliasmith, C. (2011). The neural optimal control hierarchy for motor control. *Journal of Neural Engineering, 8, 065009*. doi:10.1088/1741-2560/8/6/065009

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology, 33,* 1–40. doi:10.1016/S0065-2601(01)80003-4

Dijksterhuis, A., & Van Knippenberg, A. (2000). Behavioral indecision: Effects of self-focus on automatic behavior. *Social Cognition, 18,* 55–74. doi:10.1521/soco.2000.18.1.55

DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology, 23,* 263–287. doi:10.1146/annurev.soc.23.1.263

Eliasmith, C. (2004). Learning context sensitive logical inference in a neurobiological simulation. In S. Levy & R. Gayler (Eds.), *AAAI fall symposium: Compositional connectionism in cognitive science* (pp. 17–20). Palo Alto, CA: AAAI Press.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 1035–1054). Amsterdam, the Netherlands: Elsevier. doi:10.1016/B978-008044612-7/50102-5

Eliasmith, C. (in press). *How to build a brain*. Oxford, England: Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Erasmus University. (2012). *Report by the Committee for Inquiry Into Scientific Integrity, 1 June 2012*. Retrieved from http://www.eur.nl/fileadmin/ASSETS/press/2012/Juli/report_Committee_for_inquiry_prof._Smeesters.publicversion.28_6_2012.pdf.

Fazio, R. H., & Towles-Schwenn, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97–116). New York, NY: Guilford Press.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11,* 77–83. doi:10.1016/j.tics.2006.11.005

Fitzsimons, G. M., & Bargh, J. A. (2003). Thinking of you: Nonconscious pursuit of interpersonal goals associated with relationship partners. *Journal of Personality and Social Psychology, 84,* 148–164. doi:10.1037/0022-3514.84.1.148

Fogassi, L. (2011). The mirror neuron system: How cognitive functions emerge from motor organization. *Journal of Economic Behavior & Organization, 77,* 66–75. doi:10.1016/j.jebo.2010.04.009

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18,* 1050–1057. doi:10.1111/j.1467-9280.2007.02024.x

Förster, J., Liberman, N., & Higgins, E. T. (2005). Accessibility from active and fulfilled goals. *Journal of Experimental Social Psychology, 41,* 220–239. doi:10.1016/j.jesp.2004.06.009

Francis, C., & Heise, D. R. (2006). *Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3*. Retrieved from http://www.indiana.edu/~socpsy/ACT/interact.htm

Freud, S. (1960). Das Ich und das Es [The ego and the id]. In A. Holder (Ed.), *Das Ich und das Es und andere metapsychologische Schriften* (pp. 171–208). Frankfurt am Main, Germany: Fischer Taschenbuch. (Original work published 1923)

Gallese, V. (2009). Motor abstraction: A neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Research, 73,* 486–498. doi:10.1007/s00426-009-0232-4

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25,* 687–717. doi:10.1521/soco.2007.25.5.687

Georgopoulos, A. P., Schwartz, A., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233,* 1416–1419. doi:10.1126/science.3749885

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9,* 558–565. doi:10.3758/BF03196313

Goffman, E. (1967). *Interaction rituals. Essays on face-to-face behavior*. New York, NY: Doubleday.

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,* 1464–1480. doi:10.1037/0022-3514.74.6.1464

Gupta, V., Hanges, P. J., & Dorfman, P. W. (2002). Cultural clusters: Methodology and findings. *Journal of World Business, 37,* 11–15. doi:10.1016/S1090-9516(01)00070-0

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42,* 335–346. doi:10.1016/0167-2789(90)90087-6

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge, England: Cambridge University Press.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the brain. *Current Biology, 17,* 323–328. doi:10.1016/j.cub.2006.11.072

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology: Interdisciplinary and Applied, 21,* 107–112. doi:10.1080/00223980.1946.9917275

Heise, D. R. (1979). *Understanding events. Affect and the construction of social action*. New York, NY: Cambridge University Press.

Heise, D. R. (2007). *Expressive order. Confirming sentiments in social action*. New York, NY: Springer.

Heise, D. R. (2010). *Surveying cultures. Discovering shared conceptions and sentiments*. Hoboken, NJ: Wiley.

Heise, D. R., & Lerner, S. J. (2006). Affect control in international interactions. *Social Forces, 85,* 993–1010. doi:10.1353/sof.2007.0007

Hofmann, M. J., Kuchinke, L., Tamm, S., Võ, M. L.-H., & Jacobs, A. M. (2009). Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience, 9,* 389–397. doi:10.3758/9.4.389

Holyoak, K., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13,* 295–355. doi:10.1207/s15516709cog1303_1

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, England: Oxford University Press. doi:10.1017/S0140525X03000153

James, W. (1950). *The principles of psychology* (Vol. 1). New York, NY: Dover. (Original work published 1890)

Kintsch. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel constraint-satisfaction theory. *Psychological Review, 103,* 284–308. doi:10.1037/0033-295X.103.2.284

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by* (2nd ed.). Chicago, IL: University of Chicago Press.

LeDoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon & Schuster.

Lieberman, M. D. (2003). Reflexive and reflective judgment processes:

A social cognitive neuroscience approach. In J. P. Forgas, K. D. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44–67). Cambridge, England: Cambridge University Press.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences, 35,* 121–143. doi:10.1017/S0140525X11000446

Loersch, C., & Payne, K. B. (2011). The situated inference model: An integrative account of the effects of primes on perception, behavior, and motivation. *Perspectives on Psychological Science, 6,* 234–252. doi:10.1177/1745691611406921

MacKinnon, N. J. (1994). *Symbolic interactionism as affect control*. Albany, NY: State University of New York Press.

MacKinnon, N. J., & Heise, D. R. (2010). *Self, identity, and social institutions*. New York, NY: Palgrave Macmillan.

Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38,* 299–337. doi:10.1146/annurev.ps.38.020187.001503

McClelland, J. A., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review, 88,* 375–407. doi:10.1037/0033-295X.88.5.375

Mead, G. H. (1934). *Mind, self, and society from the standpoint of a social behaviorist*. Chicago, IL: University of Chicago Press.

Miron, M. S. (1969). What is it that is being differentiated by the semantic differential? *Journal of Personality and Social Psychology, 12,* 189–193. doi:10.1037/h0027714

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50,* 655–663. doi:10.1016/j.neuron.2006.03.040

Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The ACS (attitudes as constraint satisfaction) model. *Psychological Review, 115,* 733–759. doi:10.1037/0033-295X.115.3.733

Morgan, R., & Heise, D. R. (1988). Structure of emotions. *Social Psychology Quarterly, 51,* 19–31. doi:10.2307/2786981

Niedenthal, P. M., Winkielman, P., Mondillon, L., & Vermeulen, N. (2009). Embodiment of emotion concepts. *Journal of Personality and Social Psychology, 96,* 1120–1136. doi:10.1037/a0015574

Oatley, K. (1992). *Best laid schemes: The psychology of emotions*. Cambridge, England: Cambridge University Press.

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology, 12,* 194–199. doi:10.1037/h0027715

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana, IL: University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62,* 42–55. doi:10.1037/h0048153

Panksepp, J. (2000). Emotions as natural kinds within the mammalian brain. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 137–156). New York, NY: Guilford Press.

Parisien, C., & Thagard, P. (2008). Robosemantics: How Stanley the Volkswagen represents the world. *Minds & Machines, 18,* 169–178. doi:10.1007/s11023-008-9098-2

Peirce, C. S., & Welby-Gregory, V. (1977). *Semiotic and significs: The correspondence between Charles S. Peirce and Victoria Lady Welby* (C. S. Hardwick & J. Cook, Eds.). Bloomington, IN: Indiana University Press.

Percival, D. B. (1993). *A collection of Lisp functions to simulate stationary random processes*. Retrieved from http://lib.stat.cmu.edu/sapaclisp/random.lisp

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48,* 175–187. doi:10.1016/j.neuron.2005.09.025

Plate, T. A. (2003). *Holographic reduced representations*. Stanford, CA: CSLI.

Prinz, W. (1987). Ideo-motor action. In H. Heuer & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 47–76). Hillsdale, NJ: Erlbaum.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* 410–430. doi:10.1002/bs.3830120511

Read, S. J., & Miller, L. (1993). Rapist or regular guy? Explanatory coherence in the constuction of mental models of others. *Personality and Social Psychology Bulletin, 19,* 526–540. doi:10.1177/0146167293195005

Read, S. J., & Miller, L. (1998). *Connectionist models of social reasoning and social behavior*. Mahwah, NJ: Erlbaum.

Read, S. J., & Simon, D. (2012). Parallel constraint satisfaction as a mechanism for cognitive consistency. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 66–86). New York, NY: Guilford Press.

Romney, A. K., Boyd, J. P., Moore, C. C., Batchelder, W. H., & Brazill, T. J. (1996). Culture as shared cognitive representations. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 93,* 4699–4705. doi:10.1073/pnas.93.10.4699

Rumelhart, D. E., McClelland, J. A., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Sadler, P., Ethier, N., & Woody, E. (2011). Interpersonal complementarity. In L. M. Horowitz & S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 123–142). New York, NY: Wiley.

Scherer, K. R., Dan, E. S., & Flykt, A. (2006). What determines a feeling's position in affective space? A case for appraisal. *Cognition & Emotion, 20,* 92–113. doi:10.1080/02699930500305016

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. New York, NY: Oxford University Press.

Schneider, A. (2004). The ideal type of authority in the United States and Germany. *Sociological Perspectives, 47,* 313–327. doi:10.1525/sop.2004.47.3.313

Scholl, W. (in press). The socio-emotional basis of human interaction and communication: How we construct our social world. *Social Science Information, 52*. doi:10.1177/0539018412466607

Schröder, T. (2011). A model of language-based impression formation and attribution among Germans. *Journal of Language and Social Psychology, 30,* 82–102. doi:10.1177/0261927X10387103

Schröder, T., Netzel, J., Schermuly, C. C., & Scholl, W. (in press). Culture-constrained affective consistency of interpersonal behavior: A test of affect control theory with nonverbal expressions. *Social Psychology*. doi:10.1027/1864-9335/a000101

Schröder, T., Rogers, K. B., Ike, S., Mell, J., & Scholl, W. (2012). *Affective meanings of stereotyped social groups in cross-cultural comparison*. Manuscript submitted for publication.

Schröder, T., & Scholl, W. (2009). Affective dynamics of leadership: An experimental test of affect control theory. *Social Psychology Quarterly, 72,* 180–197. doi:10.1177/019027250907200207

Schröder, T., Stewart, T. C., & Thagard, P. (2012). *Neural dynamics of intention, emotion, and action*. Manuscript submitted for publication.

Shah, J. (2003). Automatic for the people: How representations of significant others implicitly affect goal pursuit. *Journal of Person-*

ality and Social Psychology, 84, 661–681. doi:10.1037/0022-3514.84.4.661

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLOS One, 3,* e1394. doi:10.1371/journal.pone.0001394

Shultz, T. R., & Lepper, M. R. (1998). The consonance model of dissonance reduction. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 211–244). Mahwah, NJ: Erlbaum.

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review, 6,* 283–294. doi:10.1207/S15327957PSPR0604_03

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7,* 268–288. doi:10.1016/0010-0285(75)90012-2

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York, NY: Appleton-Century-Crofts.

Skrandies, W. (1998). Evoked potential correlates of semantic meaning: A brain mapping study. *Cognitive Brain Research, 6,* 173–183. doi:10.1016/S0926-6410(97)00033-5

Smeesters, D., Yzerbyt, V. Y., Corneille, O., & Warlop, L. (2009). When do primes prime? The moderating role of the self-concept in individuals' susceptibility to priming effects on social behavior. *Journal of Experimental Social Psychology, 45,* 211–216. doi:10.1016/j.jesp.2008.09.002

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70,* 893–912. doi:10.1037/0022-3514.70.5.893

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,* 108–131. doi:10.1207/S15327957PSPR0402_01

Smith, H. W., Matsuno, T., & Ike, S. (2001). The affective basis of attributional processes among Japanese and Americans. *Social Psychology Quarterly, 64,* 180–194. doi:10.2307/3090132

Smith-Lovin, L. (1987a). The affective control of events within settings. *Journal of Mathematical Sociology, 13,* 71–101. doi:10.1080/0022250X.1987.9990027

Smith-Lovin, L. (1987b). Impressions from events. *Journal of Mathematical Sociology, 13,* 35–70. doi:10.1080/0022250X.1987.9990026

Smith-Lovin, L., & Heise, D. R. (1988). *Analyzing social interaction: Advances in affect control theory*. New York, NY: Gordon & Breach Science.

Springer, A., & Prinz, W. (2010). Action semantics modulate action prediction. *Quarterly Journal of Experimental Psychology, 63,* 2141–2158. doi:10.1080/17470211003721659

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in Decision Neuroscience, 6*. doi:10.3389/fnins.2012.00002

Stewart, T. C., & Eliasmith, C. (2012). Compositionality and biologically plausible models. In W. Hinzen, E. Machery, & M. Werning (Eds.), *Oxford handbook of compositionality*. New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199541072.013.0029

Stewart, T. C., & Eliasmith, C. (2011). Neural cognitive modeling: A biologically constrained spiking neuron model of the Tower of Hanoi task. In L. Carson, C. Haelscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society*. Wheat Ridge, CO: Cognitive Science Society.

Stock, A., & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research, 68,* 176–188. doi:10.1007/s00426-003-0154-5

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

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. (2012a). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P. (2012b). *Creative intuition: How EUREKA results from three neural mechanisms*. Manuscript submitted for publication.

Thagard, P. (2012c). Mapping minds across cultures. In R. Sun (Ed.), *Grounding social sciences in cognitive sciences* (pp. 35–62). Cambridge, MA: MIT Press.

Thagard, P. (in press). The self as a system of multilevel interacting mechanisms. *Philosophical Psychology*.

Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition, 17,* 811–834. doi:10.1016/j.concog.2007.05.014

Thagard, P., & Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram & D. B. Leake (Eds.), *Goal-driven learning* (pp. 439–454). Cambridge, MA: MIT Press.

Thagard, P., & Schröder, T. (2012). *Emotions as semantic pointers: Constructive neural mechanisms*. Manuscript submitted for publication.

Thagard, P., & Stewart, T. C. (2011). The AHA! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35,* 1–33. doi:10.1111/j.1551-6709.2010.01142.x

Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22,* 1–24. doi:10.1207/s15516709cog2201_1

Tiedens, L. Z., & Fragale, A. R. (2003). Power moves: Complementarity in dominant and submissive non-verbal behavior. *Journal of Personality and Social Psychology, 84,* 558–568. doi:10.1037/0022-3514.84.3.558

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming with to be ignored objects. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 37,* 571–590. doi:10.1080/14640748508400920

Tipper, S. P., & Weaver, B. (2008). Negative priming. *Scholarpedia, 3,* 4317. doi:10.4249/scholarpedia.4317

Todorov, A. B., Fiske, S. T., & Prentice, D. A. (2011). *Social neuroscience: Toward understanding the underpinnings of the social mind*. New York, NY: Oxford University Press.

Tsakiris, M., & Haggard, P. (2010). Neural, functional, and phenomenological signatures of intentional actions. In F. Grammont, D. Legrand, & P. Livet (Eds.), *Naturalizing intention in action* (pp. 39–64). Cambridge, MA: MIT Press.

Van Overwalle, F. (1998). Causal explanation as constraint satisfaction: A critique and feedforward alternative. *Journal of Personality and Social Psychology, 74,* 312–328. doi:10.1037/0022-3514.74.2.312

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30,* 829–858. doi:10.1002/hbm.20547

Van Overwalle, F., & Heylighen, F. (2006). Talking nets: A multiagent connectionist approach to communication and trust between individuals. *Psychological Review, 113,* 606–627. doi:10.1037/0033-295X.113.3.606

Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review, 9,* 231–274. doi:10.1207/s15327957pspr0903_3

Von der Malsburg, C. (1981). *The correlation theory of brain function*. Göttingen, Germany: Max-Planck Institute for Biophysical Chemistry. Retrieved from http://cogprints.org/1380/1/vdM_correlation.pdf

Wheeler, S. C., & DeMarree, K. G. (2009). Multiple mechanisms of prime-to-behavior effects. *Social and Personality Psychology Compass, 3,* 566–581. doi:10.1111/j.1751-9004.2009.00187.x

Wheeler, S. C., DeMarree, K. G., & Petty, R. E. (2007). Understanding the role of the self in prime-to-behavior effects: The active-self account. *Personality and Social Psychology Review, 11,* 234–261. doi:10.1177/1088868307302223

Wheeler, S. C., Jarvis, W. B. G., & Petty, R. E. (2001). Think unto others: The self-destructive impact of negative racial stereotypes.

Journal of Experimental Social Psychology, 37, 173–180. doi: 10.1006/jesp.2000.1448

Wheeler, S. C., Smeesters, D., & Kay, A. C. (2011). Culture modifies the operation of prime-to-behavior effects. *Journal of Experimental Social Psychology, 47,* 824–829. doi:10.1016/j.jesp.2011.02.018

Williams, L. E., & Bargh, J. A. (2008, October 24). Experiencing physical warmth promotes interpersonal warmth. *Science, 322,* 606–607. doi: 10.1126/science.1162548

Williams, L. E., Huang, J. Y., & Bargh, J. A. (2009). The scaffolded mind: Higher mental processes are grounded in early experience of the physical world. *European Journal of Social Psychology, 39,* 1257–1267. doi:10.1002/ejsp.665

Zemack-Rugar, Y., Bettman, J. R., & Fitzsimons, G. J. (2007). The effects of nonconsciously priming emotion concepts on behavior. *Journal of Personality and Social Psychology, 93,* 927–939. doi:10.1037/0022-3514.93.6.927

Appendix

Technical Details of the Parallel-Constraint-Satisfaction Model

The model is a modification of Kunda and Thagard's (1996) IMP (short for "IMPression formation") program written in LISP. "Observed" is a function that sets up a link between a special node providing initial activation (displayed at the bottom in Figure 2) and the "self," "prime," and "target" nodes (see Figure 2). "Associate" is another function that sets up the connections displayed in Figure 2 among all the other nodes. The "Observed" and "Associate" functions have a "Degree" parameter, supposed to model the strength of the connections. In our model of priming, "Degree" is determined by empirical evaluation-potency-activity ratings of the respective concepts, as described in the main article. To compute the exchange of activation between units during simulations, the "Degree" parameter is multiplied by IMP's default values for excitation and inhibition (.04 and –.06, respectively), in order to set the weights *w*.

To determine activation of the units, each unit is given a starting activation close to 0, except the special "Observed" unit, which is fixed at 1. Repeated cycles of updating begin. Activation is allowed to range between –1 and +1. On each cycle, the activation of a unit *j*, a_j, is updated according to the following equation:

$$a_j(t+1) = a_j(t)(1-d) + \begin{cases} net_j(\max - a_j(t)) \; if \; net_j > 0 \\ net_j(a_j(t) - \min) \; otherwise \end{cases}$$

Here, *d* is a decay parameter with default value of .05 that decrements each unit at every cycle, *min* is the minimum activation (–1), *max* is the maximum activation (+1). Based on the weight w_{ij} between each unit *i* and *j*, the net input net_j to a unit is computed by

$$net_j = \sum_i w_{ij} a_i(t).$$

Typically, the model settles after between 70 and 200 cycles, returning stable activation values for all the units. After this has happened, we interpret the activation level of the action node as corresponding to the likelihood of displaying the behavior in question.

As we explained in the main article, this procedure was repeated 1,000 times for each experimental condition we simulated. In each round of simulations, the "Degree" parameter of all the nodes was adjusted according to a random number drawn from a Gaussian distribution using the means and standard deviations from empirical evaluation-potency-activity rating studies as parameters.

Received March 27, 2012
Revision received September 10, 2012
Accepted September 11, 2012 ■