Social Cognition, Vol. 33, No. 1, 2015, pp. 20–66

A CONNECTIONIST MODEL OF SPONTANEOUS TRAIT INFERENCE AND SPONTANEOUS TRAIT TRANSFERENCE: DO THEY HAVE THE SAME UNDERLYING PROCESSES?

Diana Orghian and Leonel Garcia-Marques
University of Lisbon

James S. Uleman
New York University

Dietmar Heinke
University of Birmingham

Spontaneous trait inference (STI) and trait transference (STT) refer to the inference of personality traits from behaviors. In STI the inferred trait is attached to the actor, and in STT it is attached to a communicator. Two different explanations are currently discussed in the literature regarding their underlying processes. One claims that a single associative process is responsible for both, and the second postulates an associative process for STT and an attributional process for STI. Here we propose that a dual-processing model is not necessary to account for the empirical data regarding STI and STT. Through a simple connectionist model, based on associative learning, we simulated the four major findings that distinguish STI from STT. Suggestions are made about what kind of evidence would be necessary to consider a dualistic view, and a broader use of this approach applied to dualistic versus single processing disputes is also discussed.

This research was funded by the Leonardo da Vinci training program. We thank Philip Woodgate, Matthew Cranwell, and Saheeda Mohamed-Kaleel for their help. We also thank Kimberly Quinn, Brandon Stewart, Frederico Marques, and Ludmila Nunes for their valuable comments on our research.

Address correspondence to Diana Orghian, Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal; E-mail: diana.orghian@gmail.com, or Leonel Garcia-Marques, Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal; E-mail: garcia_marques@sapo.pt.

Inferring traits and characteristics about others' personalities is a natural way of knowing each other; a way of organizing the complexity of the social world that allows us to predict others' behaviors and achieve cognitive control over one's environment. In everyday life, we sometimes intentionally form impressions about others' personalities from their behaviors, for instance, in a job interview. But predominantly, impressions occur without any intention and awareness of making such inferences, revealing the remarkable efficiency of this ability (Todorov & Uleman, 2002, 2003, 2004). By definition (Uleman, Newman, & Moskowitz, 1996), a spontaneous trait inference (STI) occurs when a personality trait of an actor (e.g., *honest*) is inferred from his/her behavior (e.g., "*Johnny told the cashier that he received too much change*") without an explicit intention to form an impression, or to infer a personality trait about the actor. Trait inference seems to be a natural and inherent process in the comprehension of the behavior itself (Winter & Uleman, 1984). But as in all the human cognitive processes, errors can occur. One specific example is when people misattribute the inferred traits to the wrong person, a person who does not enact the behavior in the description but who tells about it. Such errors are called spontaneous trait transference (STT; Carlston, Skowronski, & Sparks, 1995; Skowronski, Carlston, Mae, & Crawford, 1998).

In our social life, besides communication about the self, we are frequently presented with communications in which informants talk about others (e.g., a reporter describing a crime; someone gossiping about a third party). Now imagine that at the first lunchtime with your new colleague Mary, she mentions that a friend of hers, Adam, "*never votes in elections.*" From this description of Adam, you can infer that he is an *apathetic* person. But surprisingly, you may also get a sense that your colleague Mary is *apathetic* too and not even realize that you've formed this impression. Spontaneous trait transference (STT) can occur toward individuals who are merely describing someone else's behavior or who are associated with that behavior due to spatial-temporal contiguity. The fact that the inferred trait can also be transferred to an inanimate object (the case of the supersticious banana; Brown & Bassili, 2002) supports the idea that STTs are caused by incidental processes. STIs and STTs are detected with a variety of memory and reaction time measures (Skowronski et al., 1998; Todorov & Uleman, 2002; Uleman, Newman, & Moskowitz, 1996) and are generally independent of awareness of making them.

The STT effect is *per se* interesting because it can have practical and observable consequences in our lives. If an informant comments about something someone else did (gossiping situation), the impression formed about the informant's personality can be affected by the traits implied in behavior he is describing. A second example is reporters who frequently describe particular kinds of events (heroic acts, criminal/aggressive acts, etc.). Also in court cases, the testimony of witnesses may imply negative traits. These traits may rub off on the witnesses and may affect the judge's or jury's impression of the witnesses, consequently influencing the outcome of the trial. The same transference may occur when someone is accidentally seen in a particular scene (e.g., someone passing by a fight on the street is later associated with the violent trait). On a general note, the STT can be seen as a special example of interferences based on second-hand information (e.g., witness

statement) and STI as inferences based on first-hand information (e.g., character of the witness). We talk about personality trait transferences in the present article, but STT may extend to transference of goals, of motivations, of needs, or of other inferences made from the informant's message (for an example see Hassin, Aarts, & Ferguson, 2005).

Besides the importance of STT on its own, the discovery of STT also influenced the investigations of STI. For instance, it has focused research on the nature of the link between the person and the trait. It has also raised the question of how STI and STT differ in terms of the underlying mechanisms that produce them. Is the same process responsible for both STI and STT or are there two distinct cognitive processes behind the two phenomena? Of course, typically this question is addressed in an empirical fashion. In contrast, the present article addresses it in a theoretical fashion, that is, proposing a connectionist model of the underlying processes. To be more precise, we will propose a model of the experiments typically conducted to address this question. This way we are able to access the validity of the theoretical claim they make (e.g., two distinct processes). Moreover and importantly, we will use this model to develop new empirical studies to advance this debate.

The question regarding the processes responsible for STI and STT has been actively debated in the literature, and two views can be clearly distinguished. According to some authors, STI and STT reflect two distinct cognitive processes (e.g., Carlston & Skowronski, 2005; Crawford, Skowronski, & Stiff, 2007). STI requires an attributional process, whereas STT is based solely on simple associative links. By contrast, Bassili (Bassili, 1976; Bassili & Smith, 1986; Brown & Bassili, 2002) advanced a single process view in which both STI and STT are based on the same simple automatic associative processes. As discussed below, there are clear differences between STI and STT, and these differences have been used to argue that there are two processes. However, an associative account, omitting the attributional one, has not been ruled out.

In fact, neither of these "processes" is well specified in this literature. "Associative processes" that link one concept (a person) to another (a trait) with a single bi-directional link seem parsimonious. All this requires is the co-occurrence of a person representation (e.g., a photo) and a trait (inferred from behavior). But this provides no account of how traits are inferred from behaviors in the first place, and ignores the evidence that these inference processes do not result from simple associative links. Inferring people's traits from behavior involves not one but three links: person-trait, behavior-trait, and person-behavior. It is known that the behavior-trait link is not symmetric; people more readily infer traits from behaviors than behaviors from traits (Maass, Colombo, Colombo, & Sherman, 2001). There is also recent evidence that the semantic links between traits and behaviors, as isolated concepts, are not merely associative but causal (Kressel, 2011; Kressel & Uleman, 2010). Much is also known about links between persons and traits, and persons and behaviors, often under the heading of stereotypes (e.g., Schneider, 2004). A person's social category affects the STIs formed from their behaviors (Wigboldus, Dijksterhuis, & Knippenberg, 2003). Taken together, treating STT (and perhaps STI, as proposed by Bassili, 1976) as the result of simple associations does not account

for the potentially different nature and roles of actor-behavior or trait-behavior links, or possible interactions among these different types of links.

The attributional account of STI is also imperfect and misleading. To begin with, the term *attribution* is an ambiguous one. Attribution can be interpreted as giving explanations or causes to behavior (and these explanations can be related with personality traits or not) or it can be interpreted as making a dispositional (trait) inference (that can be explanatory but usually is not) from behaviors (Malle, 2003).

Regardless the clarity of the attribution process, STI actually has shown some of the features of intentional attributions as described by classic attribution theories (Heider, 1958; Jones & Davis, 1965; Kelley, 1967). STIs are sensitive to behaviors' valences, that is, the attribution made from negative behaviors, which are relatively uncommon and thus diagnostic, tend to be stronger than attribution made from positive behavior. The possibility of generalization to other traits (halo effect) is another example of how comparable STI and attribution theories can be. However the catalogue of possible similarities among attributions, STI, and STT is largely unexplored.

Another important point is STI's unintentional and largely unconscious nature which contrasts with attributions occurring consciously in response to unexpected behaviors (Clary & Tesser, 1983; Hastie, 1984; Kanazawa, 1992; Lau & Russell, 1980; Pyszczynski & Greenberg, 1981; Wong & Weiner, 1981), subjective loss of control (Pittman & Pittman, 1980; Swann, Stephenson, & Pittman, 1981), personal relevance (Berscheid, Graziano, Monson, & Dermer, 1976; Harvey, Town, & Yarkin, 1981), and failure (Diener & Dweck, 1978; Wong & Weiner, 1981). The intentionality argument is not very strong, though, because it has been recently shown that causal attribution can occur in a spontaneous way (for an example see Hassin, Bargh, & Uleman, 2002). But there is another relevant evidence that has to be attended regarding the causality of traits; Kressel and Uleman's (2010) work supports the view that attribution process is not even necessary for traits to function as causes, since traits are considered causes even in isolation.

Thus, both associative and attributional processes are very incomplete in terms of their explanation of the STI and STT phenomena since they only describe the trait-person link.

Nevertheless, the dichotomy between simple associations and attributions has served as a placeholder for explaining the differences between STI and STT. And the dominant view has been the dualistic view which says that STTs are based on simple associations between persons and traits (once traits are inferred by other processes), whereas STIs reflect different, more complex and deeper (although still unintended and largely unconscious) processes, and establish properties of persons rather than mere associations with them.

Therefore, we would like to initiate a deeper discussion of this particular dichotomy between these two processes, but also a broader debate on single versus dual processing dichotomies, which is a recurrent theme in psychological science. For example, the long-standing tradition of dual-process explanations includes explicit versus implicit memory (Schacter, 1987), amodal versus modal representations (e.g., Fodor, 1975; Barsalau, 1999, respectively), direct versus indirect routes

to action selection (Yoon, Heinke, & Humphreys, 2002), etc. However, there is also a long-standing tradition of using computational models to demonstrate that these dichotomies are not necessarily true. For instance, MINERVA was proposed as a demonstration that a memory model based on exemplars could account for prototype effects without the need to postulate abstraction representations (Hintzmann, 1986; Hintzmann & Ludlum, 1980). Josefowitz, Staddon, and Cerruti (2009) presented a simple associative model (Behavioral Economic Model; BEM) that does not include metacognitive processes but that can simulate animal behavior previously taken to be diagnostic of metacognition. Seidenberg and McClelland's (1989) connectionist model demonstrated that the pronunciation of words that follow regular pronunciation rules (regular words) and words that don't follow regular rules (irregular words) can be generated with a single process.

So the present article sees itself in the tradition of these demonstrations. We present a connectionist approach to STI and STT, using a connectionist model named MATIT—Model of Associative Trait Inference and Trait Transference. The aim of this model is to simulate the four main empirical differences between STI and STT with one simple autoassociative connectionist network, based on a single underlying associative process. If our simulations are successful, they suggest that one "process" can produce the main differences found between these two phenomena. Thus, the current body of empirical data is not sufficient to support the existence of two different processes.

The application of a connectionist framework to social cognition is not new. There are connectionist models for causal attribution (Read & Montoya, 1999; Van Overwalle, 1998), cognitive dissonance (Shultz & Lepper, 1996; Van Overwalle & Jordens, 2002), and impression formation (Kashima, Woolcock, & Kashima, 2000; Van Overwalle & Labiouse, 2004). In particular, Van Overwalle and Labiouse (2004) used an autoassociative network to investigate phenomena involving primacy and recency in impression formation, the asymmetric impact of ability versus morality behaviors, memory advantages for inconsistencies, assimilation and contrast in priming, and the effect of situational constrains on trait inference. However, it is important to note that computational models are often not falsifiable as they are able to fit any data, even contradictory data (e.g., Roberts & Pashler, 2000). A classical example of this problem was highlighted by Wexler's (1987) paper on Anderson's (1976) ACT theory. Wexler (1987) showed that ACT could not only model what it is meant to model (the Sternberg result), but also its opposite. Thus, we follow Roberts and Pashler's recommendation (2000) and, after presenting our model's abilities to mimic existing evidence we will present two predictions from the model and discuss how particular experiments would have the potential to contradict these predictions, that is, falsify MATIT.

The present work used a model inspired by Van Overwalle and Labiouse (2004). We believe that our simulation model is cognitively plausible and that it accounts for the most important empirical differences found in the literature between STI and STT. More importantly, by using a relatively simple model to reproduce both STI and STT phenomena, we hope to demonstrate that it is not necessary to postulate two separate processes to account for these phenomena and the differences be-

tween them. Rather, their empirical differences may result from the same process, and also from differences in the deployment of attention within the experimental paradigms (an idea developed later on in the article).

To forestall misunderstandings, below we list all we do and do not mean to show with the implementation of MATIT and this article in general. We *do not intend to* (a) present a model that describes STI and STT phenomena in their intrinsic complexity; (b) explain all the differences between STI and STT; or (c) defend a single process view. And what we *do intend to show* in the article is that the evidence used to suggest the existence of two processes is easily reproduced by a simple and purely associative model. It is crucial to note that the MATIT model is indeed a very simple model and that therefore it can easily go wrong and be disproved. Our point is not that we are able to come up with an associative model that can explain previous results. After all, given some theoretical latitude and/or ad hockery, any type of model can simulate (mimic) any pattern of data (Anderson, 1978; Garcia-Marques & Ferreira, 2011). In that sense, finding a simulation model that simulates a data pattern is like fitting a statistical model. It will be a meaningless achievement unless the model can be falsified by plausible data (e.g., Roberts & Pashler, 2000). As we will demonstrate with MATIT, the advantage of using a simple (baseline) associative model is that even when it fits the data, it can provide clear guidelines for obtaining data that will challenge the model, that is, more diagnostic data. Such a data pattern would be diagnostic in indicating what a critical experimental design would be, that would adequately test the single versus dual process views.

This article is organized as follows: (1) a description of the problem; (2) an overview of the MATIT model, qualitatively describing the architecture and how it processes and learns information (the mathematical details can be found in the Appendix); (3) seven simulations; and (4) a general discussion and conclusions.

ASSOCIATIVE VERSUS ATTRIBUTIONAL PROCESSES

The two-process view (Carlston & Skowronski, 2005) suggests that both associative and attributional processes may come into play during the spontaneous encoding of the behavioral descriptions. Attributional processes are elaborative processes activated during the encoding of behaviors and of their actors. They involve deeper mental activity that implicates attributional (causal) knowledge and logic. They produce labelled associations between traits and persons that incorporate retrievable tags that define traits as properties of people (Johnny *is* honest) or as causes of their behaviors (Johnny returned the wallet with all its money *because* he is honest; Kressel & Uleman, 2010). Attributional processes are described by classical attributional theories (Heider, 1982; Jones & Davis, 1965; Kelley, 1973). But something different is said to occur when behaviors are presented with persons who are not the actors. Then, a simple associative process occurs.

The associative process is characterized as relatively shallow and results in generic unlabelled linkages (Carlston & Smith, 1996). It is a consequence of the spa-

tial and temporal contiguity of activated constructs (trait and person). It is insensitive to the information's diagnosticity. Contrary to the attributional processes, where "is a property of" or "is a cause of" or "is an impediment to" links occur, associative links are unlabeled (Carlston & Smith, 1996; Johnny *is associated with* the concept "honest"). Furthermore, the linkages in memory are weaker because they are established through a process that involves little elaboration.

In the single process view (Bassili, 1976; Bassili & Smith, 1986; Brown & Bassili, 2002) STI, like STT, may result from automatic associative links to traits activated during the encoding of behaviors and other stimuli such actors, communicators, bystanders, or even inanimate objects that happen to be part of the context at the moment. Note that regardless of the debate about the existence/coexistence of these two processes, it is assumed that they happen during the encoding of the behavior and of the person stimulus, not during the retrieval of this information. The single process view attempts to explain differences between STI and STT through different associative strengths between the trait and the person. What could be responsible for the different levels of associative strengths in STI and STT? We propose that these differences in associative strenght result from different "activations of the representations" of the presented stimuli (behavior, actor, and the communicator).

Note that we use this term "activation of representation" in the specific sense of the connectionist framework of our model. In this framework, representations are losely related neural activities in the brain. There can be many reasons for a representation to become more activated. These include conceptual relevance (e.g., more relevant for the task), or activation just because there is more availiable attention in one case than in others (Orghian, Gancarcyk, Garcia-Marques, & Heinke, 2014). In fact, MATIT's implementation of differential activations is based on the presumed operations of "internal" attention (e.g., Chun, Golomb, & Turk-Browne, 2011). Chun et al. (2011) distinguish two types of attention, external and internal. External attention deals with perceptual information whereas internal attention operates on internal information such as the contents of working memory, task sets, etc. In MATIT we assume that all the relevant information, for example, behavioral descriptions, actors, and bystanders, is internally represented, that is, in working memory, and that we pay more (internal) attention to the actor in the STI condition than to the communicator in the STT condition which, in turn, leads to varying levels of associative strength (see Craik & Lockhart, 1972; Naveh-Benjamin, Craik, Perretta, & Tonev, 2000, for a review of evidence on the link between attention and memory). These different levels of associative strength will result in the differences between STT and STI. Note that the influence of internal attention is not a crucial asumption in MATIT.

We chose attention because there are numerous computational models postulating that attention can be responsible for the different levels of activation (e.g., Bundesen, Habekost, & Kylingsbaek, 2005; Heinke & Backhaus, 2011; Mavritsaki, Heinke, Allen, Deco, & Humphreys, 2011). Also, evidence from electrophysiological studies (e.g., see Chelazzi, Miller, Duncan, & Desimone, 1993, for one of the first findings of this type) and neuroimaging studies (see Kastner, Pinsk, De Weerd,

Desimone, & Ungerleider, 1999, for an example of external attention, and Lepsien & Nobre, 2006, for internal attention) point in the same direction. Moreover the evidence on the relationship between attention and memory as pointed out earlier is very strong (see Craik & Lockhart, 1972; Naveh-Benjamin, Craik, Perretta, & Tonev, 2000). Furthermore as we discuss below, current evidence on STI and STT does not rule out this possibility. Finally, it is generally agreed that attention is a ubitqutious process. Hence, it would be very surprising if attention were not involved in STT/STI.

EMPIRICAL DIFFERENCES BETWEEN STI AND STT

The idea that STI could be the result of attributional processes is inspired by similarities between characteristics of STI and attributions as described in classical theories (Heider, 1982; Jones & Davis, 1965; Kelley, 1973). An example is the well-known negativity effect. In this, because negative behaviors are more noncommon and non-normative, they are more diagnostic (informative) than positive behaviors. This makes attributions from negative behaviors stronger (more likely, extreme, and confident; Reeder & Brewer, 1979). Indeed, Carlston and Skowronski (2005) demonstrate that this negativity effect exists for STI but not STT, so they concluded that STI results from attributional knowledge.

Skowronski et al. (1998) suggested that the transference phenomenon (STT) can be described as a pure associative process that results from a series of three steps (see also Mae, Carlston, & Skowronski, 1999). In the first step, the trait (e.g., *helpful*) is activated during behavior comprehension (e.g., "*Ben carried the old lady's groceries across the street*"). In the second step, the inferred trait is associated with the presented person. Finally, the association made in the previous step implicitly influences the impression of the person with whom the trait was associated. Hence this process does not reflect trait judgement or attribution, but is merely a result of the simultaneous activation of trait and person, that is, an associative process. On the other hand, STI is the result of a more elaborate attributional process that accounts for its differences from STT.

Both STI and STT are difficult to control since they occur even when the perceivers are warned of the effects and are told to avoid them (Carlston & Skowronski, 2005) or under cognitive load (Crawford, Skowronski, & Stiff, 2007).

There are important empirical differences between STI and STT that compelled some investigators to suppose that they involve different processes. The first of these differences is in the magnitude of the effect that is usually greater for STI than STT (e.g., Brown & Bassili, 2002; Goren & Todorov, 2009; Skowronski et al., 1998). For instance Skowronski et al. (1998, Study 2) obtained trait ratings of targets two days after participants merely familiarized themselves with targets' photos and descriptions of their own behaviors, or their descriptions of behaviors by (not pictured and opposite-sex) acquaintances. Effects for STIs (d = .74) were about twice those for STTs (d = .35). In this article, we will show that our connectionist model can easily mimic this and others differences between STT and STI through

variations in the associative links between faces and traits (similar to Bassili, 1976; Bassili & Smith, 1986; Brown & Bassili, 2002).

Second, several studies found that if a photo of the actor is presented next to the informant during the encoding of the behavior, the transference effect is reduced or even eliminated (Crawford, Skowronski, Stiff, & Scherer, 2007; Goren & Todorov, 2009; Todorov & Uleman, 2004).

Third, a concurrent inferential task, such as asking participants to detect whether the presented person is lying about the behavioral description (about their own behavior in the case of the actor, or about the other's behavior in the case of the informant), seems to reduce STI's magnitude whereas it has no effect on STT (Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998). The authors of these studies believe that the lie-detecting task interferes with the attributional process and not with the associative one.

Finally, several studies (Carlston & Skowronski, 2005; Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998) have shown that trait generalization or halo effects are more likely for the actor than for the non-actor. It is said that this happens because in STI, the links between persons and traits are inferential/attributional and, in accordance with attribution and implicit personality theories, allow generalization from one trait (*smart*, an intellectual trait with a positive valence) to other traits (*friendly*, a social trait with the same positive valence as *smart*).

Table 1 lists the experiments and the corresponding topics each simulation intends to replicate. This set of simulations contrasting STI and STT does not exhaustively include all published studies, but only the most important ones.

Our simulations with MATIT focus on the false recognition paradigm (Todorov & Uleman, 2002, 2003, 2004) as one of the most common and recent paradigms used to study STI and STT effects. The paradigm consists of two phases. In the study phase the participants are presented with photos of faces along with one of two types of sentences. The first type of sentence includes a trait and behavior, for example, "*Mary is so helpful that she carried an elderly lady's groceries across the street.*" The second type includes only behavior, such as, "*Mary carried an elderly lady's groceries across the street*," so that the trait *helpful* is only implied in the sentence. Participants are asked to memorize the pairs of stimuli (photo and sentence). Subsequently in the test phase, they see a series of face-trait pairs and have to indicate whether the word (trait) previously appeared in the sentence paired with that particular person. For the second type of sentence, yes responses constitute false recognitions and indicate spontaneous trait inferences at encoding, because the trait wasn't actually presented but only implied. Participants show more false recognition of traits that were originally implied in the learning phase and then tested with that same face (on "matched" trials—old pairing) than traits presented with faces originally presented with other behaviors (on "mismatched" trials—new pairing). In STI conditions, the person in the photo is said to be the actor in the behavior. In the STT conditions, the person is said to be an informant or communicator, or a bystander, or a photo randomly paired with that sentence.

TABLE 1. Overview of the Simulations

Simulation 1	
Finding	STI effect stronger than STT effect (e.g., Goren & Todorov, 2009; Skowronski et al., 1998).
Method	Higher initial input value (attention-based) for the node that represents the actor than for the node that represents the irrelevant person.
Simulation 2	
Finding	Simultaneous presentation of the actor and the irrelevant person eliminates/reduces the STT effect and has no effect on STI.
Method	Increasing even more the difference in the initial input value for the actor node and the irrelevant person node.
Simulation 3	
Finding	The lie- detection instruction reduces the STI effect at the level of the STT effect (Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998).
Method	The initial input value for the behavior node was reduced and the input for both relevant and irrelevant nodes were increased to similar values.
Simulation 4	
Finding	Trait generalization is more likely in the STI than in STT (Carlston & Skowronski, 2005; Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998; Wells et al., 2011).
Method	Additional input is given where the node representing the implied trait (critical trait) is paired with another node that represents another trait (a valence consistent trait).
Simulation 5	
Demonstration	Replication of the simulation 1 with a wider range of parameters.
Simulation 6	
Double Dissociation 1	Photo repetition is expected to increase STI and decrease STT in case the dual process view is correct.
Result	STI and STT were equally affected by the manipulation.
Simulation 7	
Double Dissociation 2	Presenting each trial twice in the task-specific learning phase is expected to lead to an increase in the STT effect and to a decrease in the STI if the dual process view is correct.
Result	The manipulation affects both STI and STT in the same manner.

Each simulation sought to replicate the result of a specific experiment that we describe before describing its implementation in the model. We used the same general model throughout, and kept the simulations as close as possible to the experimental design of the actual studies. But we made some minor simplifications to facilitate modelling and facilitate understanding of the model (e.g., fewer trials and fewer stimuli). The same experimental paradigm was modelled in all the simulations to keep experiments and simulations comparable.

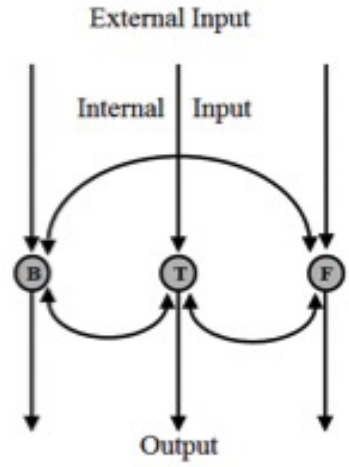

FIGURE 1. Illustration of an autoassociative recurrent network: The nodes represent the stimuli: B = behavior, T = trait, F = face. The lines represent the connections between these nodes. Each node receives internal input from other nodes which are summed to produce the internal input and also exterior activation.

THE MODEL OF ASSOCIATIVE TRAIT INFERENCE AND TRAIT TRANSFERENCE (MATIT)

The model consists of two parts: the effect of internal attention on the representation of information and the autoassociative network. We will first describe that network and then describe how we model the internal attention. Our network was inspired by the recurrent autoassociative network that was proposed by McClelland and Rumelhart (1985). However, we used a simplified version of the network as implemented by Van Overwalle and Labiouse (2004) in their work about person impression formation. As shown in Figure 1, one of the most important features of the network is that all the nodes in the system are interconnected. The network has two operation modes, a phase where the activation of the nodes is computed, and a second phase where the weights of the connections are updated. In the first phase, the model receives an external input that typically comes from the environment. Because the nodes are interconnected, the activation received from external sources spreads throughout the network. Besides the external input a node also received activation from other nodes in the network. A memory trace is created as a consequence of weight changes that are driven by the error between the internal activation generated by the network and the external input received from outside sources. The error-reduction mechanism is based on the *delta learning algorithm* (McClelland & Rumelhart, 1989) that has the function of adjusting the weights of the connections between nodes. When a node receives too much input from other nodes, this means the network is overestimating the external input of that node and the way the delta rule acts in this situations is by decreasing the weights of the connections between that node and the other nodes. In case the network underestimates the external input, the algorithm's role is to increase the weights of

the connection to better approximate the internal to the external input. The error decreases in proportion to a learning rate parameter, which determines how fast the network learns and corrects the discrepancies between the two kinds of inputs. After several external input series, the activation in the network becomes better and better in simulating/predicting the external input, and at some point it settles into a stable pattern of activations. For mathematical details see Appendix.

So, the main goal of the network and the delta algorithm is to adjust the input activation to converge on the activation received from the environment, by minimizing the difference (the error). The connection weights are initially set to zero (or random small values). Thus at the beginning of learning, these weights are small and inefficient in predicting the external input. But gradually the accuracy of the network and its ability to represent the external activation pattern increases as more external information is provided and "learned."

Each node in the network represents a construct with psychological meaning. This type of encoding is called localist, as opposed to distributed encoding (Smith & DeCoster, 1999) in which each concept is represented by a pattern of activation across a group of nodes. Distributed encoding more plausibly represents the organizations and the functioning of neurons in the brain. But localist encoding is useful when a simple demonstration is preferred. There is also some evidence (e.g., Van Overwall & Labiouse, 2004, esp. pp. 52–53) that approximately the same results that localist encoding provides can be obtained with distributed representations.

The localist representation in MATIT takes on the following form. Imagine that we present a sentence describing some behavior and two faces (the actor of the behavior and the communicator of the sentence) to a participant in the same trial for memorization. In computational terms, a way of presenting this specific trial to our algorithm is to present the following input: 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0—where the first 8 digits (1 or 0) represent 8 possible behaviors with one of them activated. The next 8 digits represent actors, and the last 8 digits represent the 8 possible bystanders. Digit one means that that specific node (and the concept it represents) is activated and zero means no activation is received from the external input for that node. So the input is a 1 × 24 matrix of zeros and ones in which the first 8 columns refer to behaviors, the second 8 to actors, and the third set of 8 columns to bystanders. To operationalize the influence of internal attention, these values (activations) are modified by taking into account two different characteristics of attention. First the item's activation should be proportional to the amount of attention that was paid to it. For instance, if an actor is strongly attended, the actor node is highly activated. This operationalization of attentional modulation is inspired by findings in electrophysiological studies (e.g., see Chelazzi, Miller, Duncan, & Desimone, 1993, for one of the first findings of this type) and neuroimaging studies (see Kastner et al., 1999, for an example of external attention, and Lepsien & Nobre, 2006, for internal attention), where the attentional state of observers was found to increase the level of activation.

Second, to mimic capacity limitations we assume that the sum of the activations of the representations is constant at one. Thus, the pattern presented earlier

is normalized turning it into (0.5 0 0 0 0 0 0 0 0.25 0 0 0 0 0 0 0 0.25 0 0 0 0 0 0 0 0). Moreover, if the activation of the actor node is increased (from 0.5 to 0.8, i.e., 0.3 more attention) the activation of other nodes has to be decreased by the same amount (0.3; less attention), the final result being: (0.8 0 0 0 0 0 0 0 0.1 0 0 0 0 0 0 0 0.1 0 0 0 0 0 0 0 0). Similar realizations of attention are found in many computational models of attention (e.g., Bundesen, Habekost, & Kylingsbaek, 2005; Heinke & Backhaus, 2011; Mavritsaki et al., 2011). Finally it is also important to note that as a result of this implementation, attention also affects the strength of the weights in associative memory. The attention in this case defines how strong is the external input that the model receives. And because the basis of the weight update algorithm between two nodes is the difference between the external (that depends closely on attention) and the internal input (that depends on the activation that comes from other nodes in the system), the associations in memory will be always affected by the attention (for more details see Appendix). This is consistent with behavioral evidence on the links between attention and memory (see Craik & Lockhart, 1972; Naveh-Benjamin, Craik, Perretta, & Tonev, 2000, further supporting this implementation of attention.

COMPUTATIONAL STUDY

For each of the simulated differences between STI and STT, there is more than one behavioral study using various methods, but we only describe and implement one of them here, the false recognition paradigm. This makes the implementation of the MATIT model easier to follow and makes the simulated effects more concrete.

GENERAL METHOD OF THE SIMULATIONS

To implement the false recognition paradigm in the model, the three types of stimuli (faces, behaviors, and traits) were represented by three nodes (see Fig. 1). The learning was based on the acquisition of patterns of weights (associations) among these nodes (see Table 2 for an example). Note that we did not model responses to trials that contained explicit traits in the behaviors. These trials were necessary for human participants so that they don't adopt a strategy of simply responding "no" on every trial. Because the model cannot adopt such a strategy (or any strategy, for that matter), these trials were not included in the simulations. Thus, behavior nodes represent sentences that contained no traits.

Each simulation consisted of two learning phases and one test phase. In the learning phases, the model was trained for two different kinds of knowledge, *world knowledge* (see Table 2) and *task-specific knowledge* (see Table 3). The world knowledge mimics the fact that spontaneous personality trait inferences rely on general knowledge about people's characteristics and their behaviors (e.g., the behavior "*shared his/her umbrella with a stranger during the rain*" with the trait *friendly*). Learning world knowledge is not part of the false recognition paradigm but is

specific to the simulation because human participants already know it, and this knowledge permits the inference of traits from behavior descriptions.

The second learning phase trained the network with the task-specific knowledge (the first step in the false-recognition paradigm), where a specific behavior (e.g., "*Anna shared her umbrella with a stranger during the rain*") is associated with a specific person (e.g., Anna's face photo). So, the learning of associations between specific traits and specific behaviors corresponds to the world knowledge, and the task-specific knowledge corresponds to the association between behaviors and specific faces.

Knowledge acquisition in the model is determined by the learning rate parameter (ε in the Appendix) and the given input (external input). The input was not always set at the same values in all the simulations because in the original experiments, the instructions, the manipulated variables, and the presumed attention differed. However the learning parameters were chosen in the first simulation and then the same values were used in all following simulations.

In order to "teach" the model the world knowledge, we had it learn the associations among a series of behavior-trait pairs (see Table 2 for specific values). Each row in Table 2 corresponds to a trial where the values represent the activation of each node. So in each trial, two nodes, a trait and a behavior, were activated simultaneously, which made the model learn that these two are associated. The learning of associations between specific trait-behavior pairs depends on the number of times this trial is given as input to the network (apart from the learning rate). The presentation frequency of each pair captures the way we learn in real life. An equivalent to the world knowledge learning would be children's learning about how to categorize behavior, when for instance the child observes someone performing a certain behavior and next hears the adult categorizing the observed behavior by naming the correspondent trait (e.g., *unfriendly*). The more often the individual is exposed to this same situation, the more he/she will think these two are related (the trait and that type of behavior). Of course this learning process constitutes a great simplification of the way children learn in real life. But we don't intend to explore the complexity of this process in the present article.

As noted above, the learning in MATIT also depends on the learning rate parameter that governs the speed of learning. The learning of the world knowledge is expected to be slow, since it is acquired over the long term, based on frequencies of exposure rather than explicit propositional learning (e.g., Gawronski & Bodenhausen, 2006).

The second part of the learning phase (see Table 3) is specifically related to the false recognition task described above—memorize pairs of behavioral sentences and faces. The input to the model is a pattern where nodes representing behaviors and nodes representing persons are activated simultaneously. Note that in both learning phases, only one behavior is associated with each trait and only one behavior with each face, and the behavior in these two learning parts is the same. In this phase, each trial (rows in Table 3) was presented only once (frequency 1), as in the experimental studies being simulated (participants saw each trial once). To make the simulations more realistic, we introduced some noise (equivalent to the

variability in behavioral data), and added a random value ranging between 0 and 0.1 to the default starting weights of zero. In both learning phases, the presentation of the inputs was randomized.

As shown in Tables 2 and 3, the activations of all nodes in the input pattern add up to 1. An input pattern refers to all nodes in each row, and each row represents a trial. This is a reflection of how MATIT models the operation of attention as explained earlier. We assume here that by asking participants to memorize several simultaneously presented stimuli, the attention (activation) will be divided between the elements to be processed. Due to the capacity limitation of human processing we kept the sum of the activation at 1 at all conditions. In the world knowledge phase the elements are the trait node and the behavior node whereas in the task-specific phase there are three different types of stimuli in play, the behavior, the actor, and the non-actor person. After learning the association between nodes, the test phase (see Table 4) was run in which we turned on some specific nodes (providing "incomplete patterns") and observed the output (the question marks in the table). The resulting output represents the completion of the patterns, that is, the activations of other nodes that were encoded in the learning phase but are not presented in the input for test phase. Because all the nodes in the system are interconnected (differing in the weights of their connections), if we give the model patterns with some nodes activated, it will "recall" associated nodes in the network due to the spread of activation.

For each simulation, the network was run 50 times, simulating 50 participants, and within each learning phase (world knowledge and task-specific knowledge) the trials were randomized for each participant. See Appendix for a walk-through example.

SIMULATION 1—STT VERSUS STI

The first simulation models the first study from Goren and Todorov (2009). In order to investigate STI and STT effects, participants in their experiment were told that some of the sentences describe the behavior of the person presented with the sentence (actor or STI trials in which sentences were presented in blue), and that other behaviors were said to be randomly assigned to the faces (non-actor or STT trials in which sentences were presented in red). A randomly assigned irrelevant face was used rather than a "communicator" because in the communicator case, participants might infer that the communicator shares traits of those he is describing. This non-actor condition eliminates any logical association between the trait and the irrelevant person. Thus any association must reflect only simple associative processing, as in STT.

The main analyses consisted of comparing the differences in "false recognition" (activation of the behavior nodes in the simulation) on the two pairings, that is, differences between old and new trials—across each level of the relevance factor (actor and non-actor). We predicted that the difference between old and new tri-

TABLE 2. Learning Pattern: World Knowledge

Behaviors								Traits								Presented person								Non-presented person			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4
0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0

Note. In the table each row represents a trial, each block of 8 nodes represents a specific type of stimulus (behavior, trait, presented person, and non-presented person). Each block has 8 different nodes representing different stimuli of that type. Zero value means that the stimuli/node is not presented/activated in the trial/input. The activation of a specific node may vary between 0 and 1, with the constraint that the sum of activations in each trial equals 1.

als (the pairing factor) would be greater on actor trials (STI) than non-actor trials (STT), thus simulating a stronger STI than STT effect.

METHOD

In this simulation, all the trials from the world knowledge pattern had a frequency of 8, which means they were given as input to the model 8 times. As shown in Table 2, this input creates the association of 8 different behavior-trait pairs. The activation of all the nodes in the patterns sum to 1. Thus in this phase the activation for the trait node was 0.5, and for the behavior node it was 0.5 as well. The learning rate parameter (ε) for the world knowledge patterns was 0.1. The selection of this value was based on the assumption that the learning of the world knowledge is slow (which is why the value is low) and is frequency-based (which is why the frequency is 8). Note that these values for the first simulation are used in all the remaining simulations.

In order to simulate the difference between STI and STT, in the second part (the task-specific phase in the model and the first part of the false recognition paradigm), the activations of the person nodes for actor and non-actor conditions were different (see Table 3). To simulate an STI trial—where the actor of the behavior is presented in the photo—we set the activation of the behavioral node to 0.5 and the activation for the actor node to 0.5 as well.

To simulate the STT trial—where the pairing of photo and behavior is arbitrary—the input activation to the behavioral node was the same as in the STI condition (0.5) as there are no reasons to expect difference between the activation of the sentence in STI and STT. However the activation for the photo-person node was weaker (0.25) because we assume less attention was paid to this non-actor because he/she is a third and a less relevant element present in this type of trial. There is the actor mentioned in the sentence but not pictured, the behavior in the sentence, and the non-actor photo. So attention is divided among these three elements rather than two, which means less attention is paid to some of them. We set the activation for both the actor and the pictured non-actor at 0.25. As proposed by Brown and Bassili (2002), these attentional differences cause differences in the strength of associations between the person and the trait. In the simulated experiment, this difference in attention is presumably produced by instructions to participants at the beginning of the study, with blue sentences signalling STI trials and red sentences signalling STT conditions).

In Table 3, the STT trials are listed in the 1st, 2nd, 5th, and 6th rows (with 0.25 for face input activation for the presented person columns—irrelevant person—and for the non-presented person columns—the relevant one) and the remaining rows correspond to the STI trials (with 0.5 for face input activation just for the presented person that is the relevant one in these trials). Thus, we had 4 trials per condition, but half of them were used as controls (new pair trials in the FRP described before) in the test phase. The learning rate (ε) for this second learning part was higher than the learning rate for world knowledge because participants are explicitly asked to

TABLE 3. Learning Pattern: Task-Specific Knowledge

Behaviors								Traits								Presented person								Non-presented person			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4
0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0	0.25	0	0	0
0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0	0.25	0	0
0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0.25	0
0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0.25
0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0
0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0

memorize the information, which makes them expend more effort in the task so that learning is more efficient. Thus the value of the learning rate was set to 0.4.

After the model memorized the association between behaviors and faces, we examined what the model learned and compared the results with those from Goren and Todorov (2009). As in that paper, where participants were asked whether the trait was or was not presented in the sentence with the specific face, requiring them to try to recall the behavioral description, we gave the model an incomplete pattern as input (Table 4). Only faces and traits were activated. Then we observed the model's response, that is, its output in the activation of the behavior nodes.

As noted above, half the trials were new pairs (the last four rows in Table 4, two for STI and two for STT), where the traits were presented with mismatched faces. Thus, as in Goren and Todorov (2009), the overall design is a 2 (Pairing: old versus new pairing) × 2 (Relevance: actor versus non-actor) within-subjects ANOVA.

SIMULATION RESULTS AND DISCUSSION

The dependent variable in all the simulations is the final activation of the behavior node. There was a main effect of Relevance (actor versus non-actor), $F(1, 192) = 55.48$, $p < 0.001$, $\eta^2 = 0.09$, and a main effect of Pairing (old versus new pairs), $F(1, 192) = 247.48$, $p < 0.001$, $\eta^2 = 0.50$ in this simulation. There was also a significant Relevance × Pairing interaction, $F(1, 192) = 33.45$, $p < 0.001$, $\eta^2 = 0.09$ (see Figure 2). There was an STI effect because the mean activation for the old trials was significantly greater ($M = 0.16$, $SD = 0.02$) than the new trials ($M = 0.11$, $SD = 0.02$), $t(49) = 15.62$, $p < 0.001$, $d = 3.14$, 95% CI [0.05, 0.06]. The same effect occurred for the mean difference in the non-actor condition, with old trials ($M = 0.13$, $SD = 0.02$) showing more activation than new trials ($M = 0.11$, $SD = 0.02$), $t(49) = 5.40$, $p < 0.001$, $d = 1.17$, 95% CI [0.01, 0.03]. To test whether there was a significant difference between STI and STT effects, as in the empirical study we simulated, we calculated the differences between the old and new pairings for the actor and for the non-actor conditions. This difference was larger for actor ($M = 0.05$, $SD = 0.02$) than for non-actor ($M = 0.02$, $SD = 0.03$), $t(49) = 5.78$, $p < 0.001$, $d = 1.26$, 95% CI [0.02, 0.04]. The pattern of behavioral activation is shown in Figure 2.

To understand the results it is important to realize that the behavioral activation in the old pairing is the outcome of two different associations, the trait-behavior associations and the behavior-face associations. The trait-behavior associations do not vary between STI and STT conditions since they were learned in the world knowledge phase, which is identical for STI and STT. This means that if we only presented the trait in the test phase (without the face), the activation of the behavioral node would be similar in STI and STT conditions. The second type of association, behavior-face, was created in the task-specific phase where two kind of trials were used: a behavior-actor pairing that results in a strong weight between these two nodes, and a behavior–non-actor pairing where the link is weaker due to the assumed distribution of attention mentioned above. This means that in the test phase, links to the faces are responsible for the different results in actor and non-

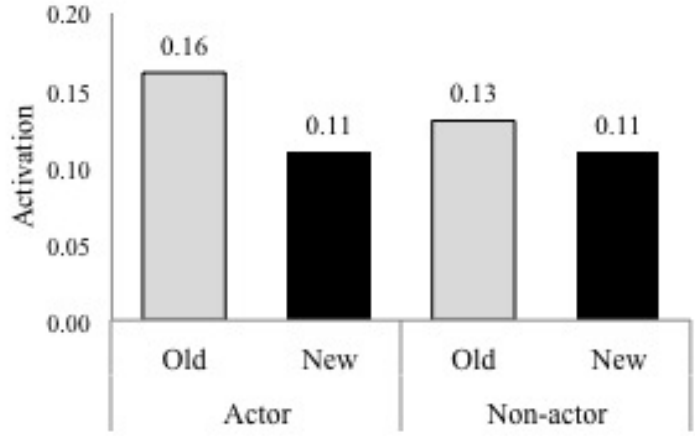

FIGURE 2. Simulation 1: Mean activation of the behavior node.

actor conditions. In the STI condition, the actors are a better cue for the retrieval of the behavior than the non-actors in the STT condition. In the STI condition, the traits and especially the faces activated the behavioral node more than in the STT condition, where the trait cues work in a similar way but the face cue is less effective, so the spread of activation to the behavioral node is less. To sum up, the results for the old parings were obtained because (1) learning world knowledge set up behavior-trait links, (2) learning the task-specific knowledge set up person-behavior links, and (3) these links were stronger for the actor than for the non-actor. For the new pairings, the activation of behaviors is the smallest (black bars in Figure 2) because these faces were not linked to the behaviors at all.

This simulation shows that this autoassociative model is able to simulate both STI and STT effects, as well as the typical magnitude difference between the two. It also suggests that it is not necessary to consider an additional process (Goren & Todorov, 2009; Skowronski et al., 1998) to explain this specific difference because differential activation of the person nodes (with stronger activation for actor than non-actor nodes) was sufficient to simulate the experimental results.

SIMULATION 2—RELEVANT AND IRRELEVANT TARGETS SIMULTANEOUSLY PRESENTED

Todorov and Uleman (2004) used a different kind of manipulation in order to eliminate or reduce the STT effect in the false recognition paradigm. In their version, both a relevant and an irrelevant target face were presented in the same trial with the behavior. The presence of relevant faces diminished the STT effect. Todorov and Uleman argued that the presence of the actor leads to a deactivation of the associative process and hence to a failure of STT (Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Leonards, 2008; Todorov & Uleman, 2004).

TABLE 4. Test Pattern

Behaviors								Traits								Presented person								Non-presented person			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4
?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.50	0	0	0	0	0	0	0	0	0	0	0
0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0
0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0
0	0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
0	0	0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0
0	0	0	0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0
0	0	0	0	0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0.5	0	0	0	0
0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0.5	0	0	0	0	0

The design of this simulation was a 2 (Faces presentation: simultaneous presentation, i.e., actor *and* non-actor, versus standard, i.e., only actor *or* only non-actor) × 2 (Pairing: old versus new pairs) × 2 (Relevance: actor versus non-actor) ANOVA, with the first factor between-Ss and the rest within-Ss.

METHOD

The world knowledge was the same as in Simulation 1. Table 5 lists the modified task-specific knowledge patterns. As in Simulation 1, actor faces were presented with a higher value (0.4 in this case) than non-actor faces (0.1). However and importantly, these two levels of attention were entered in the same rows/input reflecting the simultaneous presentation of the two faces in Goren and Todorov's experiment. Here 0.5 units of attention were divided between the two photos, and since the relevance of the actor is higher in this context (compared to a non-relevant person presented alone), more attention is paid to the actor than to the non-actor. The remaining 0.5 was assigned to the behavior. This realization of Goren and Todorov's experiment illustrates how our idea of divided attention can easily be generalized to a different experimental design. In the test pattern as in all other simulations, only one face (the actor's or non-actor's) and one trait was activated, in the incomplete pattern to be completed by spreading activation.

SIMULATION RESULTS AND DISCUSSION

We found main effects of face Presentation, $F(1, 388) = 29.16$, $p < 0.001$, $\eta^2 = 0.04$; of Pairing, $F(1, 388) = 302.87$, $p < 0.001$, $\eta^2 = 0.30$; and of Relevance, $F(1, 388) = 169.39$, $p < 0.001$, $\eta^2 = 0.12$. The was no significant Presentation × Relevance effect, $F(1, 388) = 1.46$, $p = 0.23$, $\eta^2 = 0.00$, but importantly there was a significant Presentation × Pairing effect, $F(1, 388) = 27.94$, $p < 0.001$, $\eta^2 = 0.03$, and also a significant Relevance × Pairing interaction, $F(1, 388) = 74.88$, $p < 0.001$, $\eta^2 = 0.10$. There was no 3-way interaction, $F(1, 388) = 0.05$, $p = 0.82$, $\eta^2 = 0.00$. But note that this is not the main result we are looking for, because we are interested in comparing very specific cells of this design (specifically the simultaneous presentation condition).

When targets were presented alone, as in a standard STI/STT study, both effects emerged. The difference in behavior activation between old and new trials was significant for the actor, $t(49) = 15.62$, $p < 0.001$, $d = 3.14$, 95% CI [0.05, 0.06], and was smaller but still significant for non-actor, $t(49) = 5.40$, $p < 0.001$, $d = 1.17$, 95% CI [0.01, 0.03]. STI persisted when targets were presented simultaneously, with a new versus old trials mean difference of 0.04 ($SD = 0.02$), $t(49) = 11.38$, $p < 0.001$, $d = 2.09$, 95% CI [0.03, 0.04]. But consistent with past empirical findings, STT disappeared. The difference in activation between old ($M = 0.11$, $SD = 0.02$) and new trials ($M = 0.11$, $SD = 0.02$) for the non-actor showed no significant STT effect, $t(49) = 0.71$, $p = 0.48$, $d = 0.14$, 95% CI [0.00, 0.01], in the simultaneous face presentation condition.

TABLE 5. Learning Pattern: Task-Specific Knowledge

Behaviors								Traits								Presented person								Non-presented person							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0
0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0
0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0	0	0
0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0	0
0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0	0
0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1	0
0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0.1

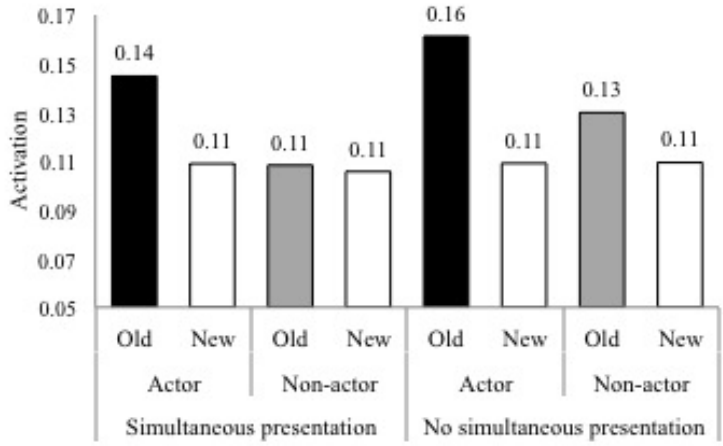

FIGURE 3. Simulation 2: The results of Simultaneous Presentation versus No Simultaneous Presentation simulation; the dependent variable is the behavioral activation difference between old and new pairing conditions.

Thus, the simultaneous presentation of the relevant and irrelevant targets affected STT but not STI (Fig. 3). In the model this occurs because of the different values in the learning input (different levels of attention), without any assumption about differences in the processes as posited by some authors (Crawford, Skowronski, & Stiff, 2007; Crawford et al., 2008; Goren & Todorov, 2009, Study 4; Todorov & Uleman, 2004).

SIMULATION 3—LIE-DETECTION TASK

Crawford and collaborators, in 2007, conducted a study where the savings-in-relearning paradigm was combined with a lie-detection task. As in previous savings-in-relearning experiments, participants were first exposed to photos paired with descriptions of behaviors. However, rather than memorizing or familiarizing themselves with the stimuli, Crawford, Skowronki, Stiff, and Scherer (2007) asked participants to decide whether the person in each photo was lying. Relevance was manipulated with self-descriptive relevant actors versus other-informant irrelevant persons. After the lie-detection task, the same photos were presented again but this time paired with a single word. The task here was to memorize the word-photo pairs. In some of these trials, photos previously seen were paired with traits implied by behaviors presented in the initial phase with that picture; others had novel photo-trait pairs. Finally in the last part, participants had to recall the words/traits (from the memorization task) that were cued with actor, informant, or novel photos. Recall performance showed the saving effect, that is, photo-word pairs that repeated photo-trait pairs from the behavioral presentation task were learned better than novel pairs.

The lie-detection manipulation resulted in diminishing the recall performance in the self-descriptive condition (actor) leading to similar sized STI and STT effects. Crawford, Skowronski, Stiff, and Scherer (2007) concluded from these results that the attributional process that normally causes strong actor-trait linkages (in the initial familiarization or memory tasks) is disrupted by the lie-detection instruction, leaving only association processes common to STI and STT in place. Despite this explanation, it remains unclear exactly what happens when we ask a participant to detect whether the person in the photo is lying. Here we assume that rather than disrupting attributional processes, more attention is paid to the photos under lie detection, and that the amount of attention is similar in STI and STT conditions. The rationale behind these assumptions is that by asking participants to detect whether the person is lying, most of their attention focuses on the face in the photo, regardless the relevance (whether the person is communicating her own behavior or someone else's behavior). This assumption is supported by the lie-detection literature which is replete with evidence that people believe that they can detect lying by carefully attending to the actor's appearance and facial behavior—averted gaze, speech fluency, etc.—even though most evidence contradicts this (e.g., Bond & DePaulo, 2006). The behavioral nodes receive less activation, but the same amount in STI and STT, because people are not even asked to memorize the behaviors. (The instruction is to detect liars.)

The design of this simulation is a 2 (Instruction: lie detection versus memorization) × 2 (Pairing: old versus new pairs) × 2 (Relevance: actor versus non-actor) ANOVA, the first factor being between-Ss and the rest within-Ss. Compared with previous simulations, here instructions vary between-Ss and these produce different activation values.

METHOD

To simulate this result, the learning rate, the simulated number of subjects in each instruction condition, and the world knowledge inputs were kept the same as in the first simulation. The main difference was in the activation values for both actor (STI) and non-actor (STT) faces, reflecting our assumptions about the distribution of attention in the lie-detection condition (see Table 6). The activation for the behavior was only 0.1 in this simulation because it was not so important in this specific task where the participant's focus is mainly on the person in the picture. The actor's face was set to 0.9 and the non-actor's face was set to a similar value (0.85), while the remaining 0.05 was reserved for the actor mentioned in the STT sentences. The nodes for the actor and non-actor faces were both highly activated because our hypothesis is that participants strongly focused on the photos to detect whether or not the persons were lying.

SIMULATION RESULTS AND DISCUSSION

The ANOVA on behavior activation showed a main effect for Pairing type, $F(1, 388) = 193.40$, $p < 0.001$, $\eta^2 = 0.33$; for Relevance, $F(1, 388) = 30.55$, $p < 0.001$, $\eta^2 = 0.04$; and for Instruction (lie detection versus memorization), $F(1, 388) = 12.70$, $p < 0.001$, $\eta^2 = 0.02$. We also found a Pairing × Relevance interaction, $F(1, 388) = 16.09$, $p > 0.001$, $\eta^2 = 0.02$; a Relevance × Instruction interaction, $F(1, 388) = 12.93$, $p < 0.001$, $\eta^2 = 0.02$; a Pairing × Instruction interaction, $F(1, 388) = 26.01$, $p < 0.001$, $\eta^2 = 0.04$; and a Pairing × Relevance × Instruction interaction, $F(1, 388) = 22.26$, $p < 0.001$, $\eta^2 = 0.03$. Breaking this down by instruction, the difference between old and new pairs under lie-detection instructions showed an STI effect ($M = 0.02$, $SD = 0.03$) but not any larger than the STT effect ($M = 0.02$, $SD = 0.03$), $t(49) = -0.31$, $p = 0.76$, $d = -0.06$, 95% CI [-0.01, 0.01] (see Fig. 4). STI was greater under memorization than lie detection, $t(49) = 6.35$, $p < 0.001$, $d = 1.34$, 95% CI [0.02, 0.05], whereas STT was not, $t(49) = 0.40$, $p = 0.69$, $d = 0.07$, 95% CI [-0.01, 0.01].

These results demonstrate that our assumptions about attention allocation and our corresponding implementation of the lie-detection task in the model successfully replicated the experimental data. Changing the relative activation of faces and behaviors made it possible to replicate the behavioral data. With the lie-detection instruction, both types of faces receive roughly the same amount of attention which produced similar effects for STI and STT.

SIMULATION 4—GENERALIZATION EFFECT

Additional support for the two-process view comes from evidence that the behavior-trait inferences generalize to other traits in STI conditions, whereas no such halo effects can be found in STT conditions (Carlston & Skowronski, 2005; Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998). This finding is often interpreted as evidence that attributional processes entail implicit theories of personality and thus generalize to other trait dimensions.

Experimentally, the halo effect is usually investigated using a trait rating task (e.g., Carlston & Skowronski, 2005). The first part of the experiment is similar to the first part of the false recognition paradigm. In the second part of the study, participants are asked to rate how much of a specific trait each person possess. Three different types of traits are used: a critical trait that was implied by the behavior in the first part of the study (e.g., *helpful*), a trait consistent with the critical trait's evaluative valence (e.g., *smart*), and a trait inconsistent with the critical trait's valence (e.g., *rude*). In the STI condition, the ratings for the congruent traits were significantly above chance and higher than in the STT condition, where they were not above chance. The facts that the transference effect was specific to traits implied by the informants' descriptions, and that the impression of the actors was influenced by the implied traits' valence activating non-implied traits with the same valence, was taken as evidence of attributional processes in STI (Carlston & Skowronski,

TABLE 6. Learning Pattern: Task-Specific Knowledge

Behaviors								Traits								Presented person								Non-presented person							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0	0	0.05	0	0	0	0	0	0	0
0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0	0	0.05	0	0	0	0	0	0
0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0	0	0.05	0	0	0
0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0	0	0.05	0	0
0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0

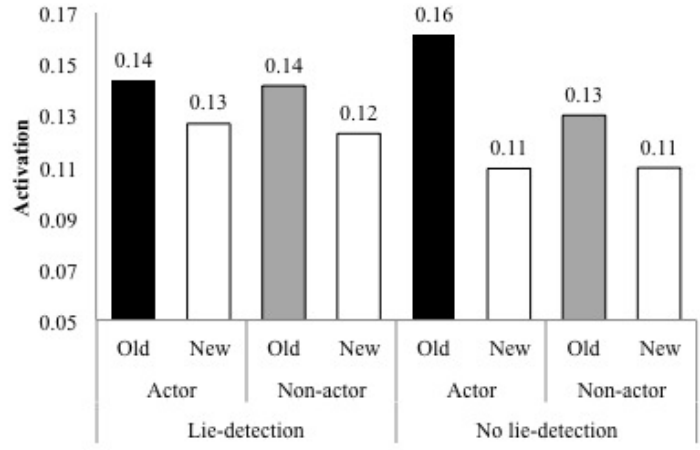

FIGURE 4. Simulation 3: The results of Lie-Detection versus No Lie-Detection simulation; the dependent variable is the activation difference between the old and new pairing conditions.

2005; Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998).

From the simulations so far, it may not be immediately obvious how MATIT could mimic these findings. The success of the simulations is based on the idea that the distribution of attention affects the strength of the weights in the autoassociative memory. However, it is difficult to see how the halo effect could be based on such a relationship. Indeed, simulating the halo effect is based on the autoassociative memory only, not on attention allocations. Thus as part of the MATIT's world knowledge, valence-consistent traits can be associated with each other just as traits can be linked with trait-implying behaviors. In other words, the generalization effect can be realized through MATIT's autoassociative memory. Such an explanation of the halo effect would be consistent with classic models of implicit personality theory that assume inter-trait relations (e.g., Anderson, 1995; Carlston & Skowronski, 2005; Schneider, 1973).

The essential design of this study is a 2 (Pairing: old versus new pairs) × 2 (Relevance: actor versus non-actor) × 2 (Trait: implied versus valence-consistent) ANOVA, all within-Ss.

METHOD

In this simulation as in the last, we focused on the important conditions and designed a slightly simplified version of the original experimental procedure. We only used implied traits and valence-consistent traits, since the valence-inconsistent traits are not crucial for our demonstration. Besides, the valence-inconsistent trait results did not vary with the actor/non-actor manipulation (Carlston & Skowronski, 2005).

In order to simulate implied and valence-consistent traits in world knowledge, we used two different frequencies for the trait-behavior pairs. For the implied traits (see the first 8 rows in the Table 7) the world knowledge patterns were exactly the same as in Simulation 1 and the frequency was also the same (8). For the valence-consistent traits, the pattern was trained with frequency 2, where the implied traits were associated with the valence-consistent traits, equivalent in real life to fewer observations of the implied traits co-occurring with valence-consistent ones (see from 9th to 16th row in Table 7). The learning rate for world knowledge was 0.1.

In the task-specific learning, the pattern was similar to Simulation 1, where the activation of the actor node (0.5) is larger than the activation of the non-actor node (0.25 plus 0.25 for the actor in the sentence). The learning rate in the task-specific knowledge was 0.4. The test pattern was similar to Simulation 1 as well; where pairs of faces and traits (some of the traits were "implied," i.e., were directly associated to behaviors and other traits were "valence-consistent," i.e., were directly associated to other "implied" traits and not to behaviors) were presented, and the activation of behavioral nodes was analyzed.

SIMULATION RESULTS AND DISCUSSION

Figure 5 shows the mean activations of the behavior nodes. A repeated measures 2 (Pairing: old versus new pairings) × 2 (Relevance: actor versus non-actor) × 2 (Trait: implied versus valence-consistent) within-Ss ANOVA found a main effect of Pairing, $F(1, 384) = 246.34$, $p < 0.001$, $\eta^2 = 0.10$; a main effect of Relevance, $F(1, 384) = 30.48$, $p < 0.001$, $\eta^2 = 0.01$; and a main effect of Trait, $F(1, 384) = 2065.16$, $p < 0.001$, $\eta^2 = 0.72$. This last effect shows that the activation for implied traits ($M = 0.13$, $SD = 0.03$) exceeded that for valence-consistent ones ($M = 0.02$, $SD = 0.03$).

Of the interaction effects, only the Pairing × Relevance interaction was significant, $F(1, 384) = 81.67$, $p < 0.001$, $\eta^2 = 0.02$, replicating the Simulation 1 results. There was no significant 3-way interaction, $F(1, 384) = 0.05$, $p = 0.83$, $\eta^2 = 0.00$, but again it is not this interaction we are looking for. The Pairing × Relevance interaction occurred because differences between old and new trials were greater for the actor ($M = 0.05$) than for the non-actor ($M = 0.02$), with $t(49) = 4.92$, $p < 0.001$, $d = 1.02$, 95% CI [0.02, 0.05] for the case of the implied traits. This was equally true for valence-consistent traits, with $t(49) = 6.29$, $p < 0.001$, $d = 1.10$, 95% CI [0.02, 0.05], showing that the difference between old and new trials for the actor, $M = 0.06$, was greater than the difference for non-actor, $M = 0.02$, that is, halo effect for the actor was stronger than the halo effect for the non-actor. Thus, by looking at the valence-consistent traits we can conclude that there is a halo effect for STIs, $t(49) = 12.65$, $p < 0.001$, $d = 2.66$, 95% CI [0.05, 0.07], and a smaller halo effect for STT, $t(49) = 5.30$, $p < 0.001$, $d = 1.02$, 95% CI [0.02, 0.03].

The behavioral activations in Figure 5 can be understood as resulting from: (1) behavior-trait links in world knowledge; (2) behavior-face links from the learning trials in which (3) actor faces are linked more strongly than non-actor faces, as in Simulation 1; and (4) indirect behavior links with valence-consistent traits through

TABLE 7. Learning Pattern: World Knowledge

Behaviors								Implied traits								Valence-consistent traits								Presented person								Non-presented person			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4
0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0

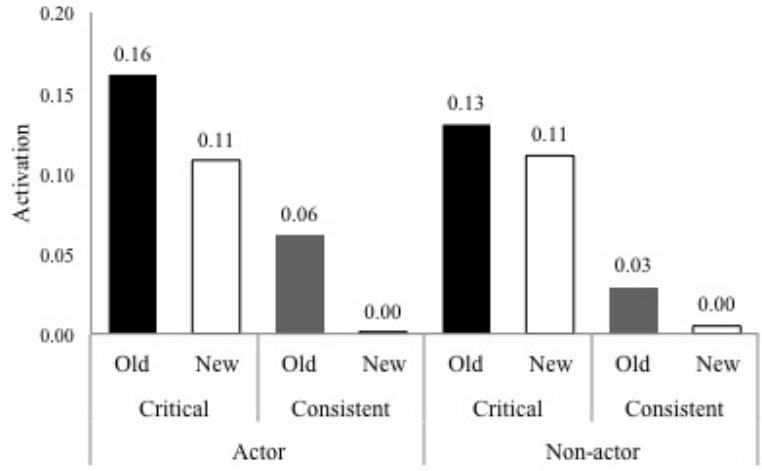

FIGURE 5. Simulation 4: Mean activation of the behavior node.

their links with implied traits. The pairs of faces and traits presented at test determine which of these links is activated and hence the net levels of behavioral activation shown in Figure 5.

SIMULATION 5: THE ROBUSTNESS OF THE MODEL

In this section we examine the robustness of the model in terms of the parameters. The robustness of the model can be tested by varying the parameters and testing whether the different parameter settings reproduce the empirical results. Since the parameters of the first simulation form the basis for all subsequent simulations, we used this for initial parameters. We also focused on the two crucial parameters, differences of attention in the STT condition and the STI condition, and the learning rate. The former is crucial because if it is very small there would not be a difference between STI and STT (darker pentagon markers in Figure 6). On the other hand, if too large, that is, much more attention is paid to the actor than to the non-actor, the STT effect is not observed anymore (square markers in Figure 6). Furthermore, if the learning rate is too small no learning would occur. Figure 6 shows the results.

The attentional difference ranged from 0 to 0.5 (with a 0.05 interval). Note that the remaining 0.5 is assigned to the behavior node. The learning rate ranged from 0 to 1 (with a 0.1 interval). Thus we run the simulation 230 times using the method described in the Simulation 1 section. The lighter pentagon markers in Figure 6 indicate that MATIT can replicate the results of Experiment 1 across a broad range of parameter settings, that is, our model is very robust. Similar explorations of robustness could be done for the other results modelled here, but would take us well beyond the scope of this article.

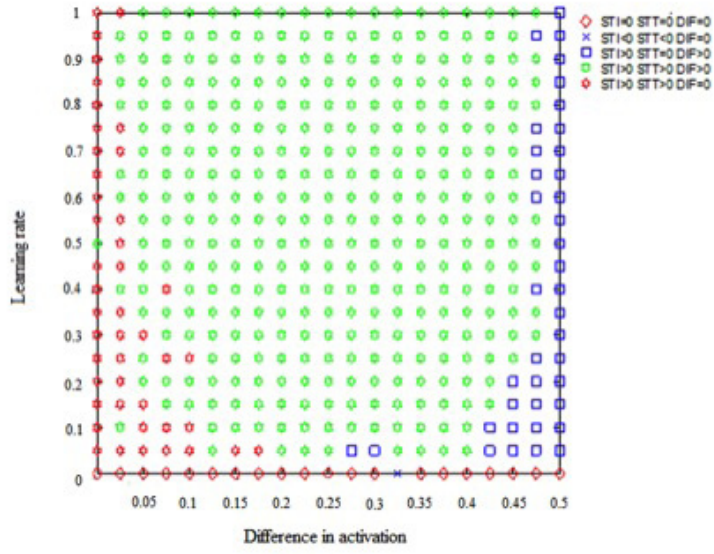

FIGURE 6. Simulation 5: The graph illustrates the robustness of the model in terms of the two parameters, the actor minus non-actor difference in activation (attention) on the x-axis and learning rate (y-axis). The lighter pentagons indicate the parameter settings for which the simulation replicated the results from Simulation 1 (STI > 0, STT > 0, STI > STT). The diamond markers indicate that MATIT did not produce a significant effect. For the darker pentagons, MATIT produced STT and STI effect, but there was no difference between them (STI > 0, STT > 0, DIFF = 0). The squares indicate an STI effect but no STT effect.

SIMULATION 6: FALSIFICATION—DOUBLE DISSOCIATION 1

This simulation is the first of two simulations which examine how MATIT might be falsified. The previous simulation results suggest that MATIT could be falsified by a double dissociation (e.g., Dunn & Kirsner, 2003) as all simulated effects constitute single dissociations. In the context of the STT/STI-effects, a double dissociation would occur if one effect (e.g., STI) increases while the other effect (e.g., STT) decreases (i.e., shows a worse performance than baseline). Our parameter scan did not show such an effect. That is, any combination of the distribution of attention and learning rate could not simulate the experimental result of a double dissociation. (Even though this result is based on the particular method of Simulation 1, and Simulation 4 showed it is possible to model other methods in MATIT, we focus on this simulation method for now.) When considering the question of falsification, it is also important that the falsification is achieved by plausible data (e.g., Roberts & Pashler, 2000). Hence the question, is a double dissociation a potential outcome of an experiment?

Interestingly, a recent study by Carlston and Skowronski (2005) appeared to find such an effect. Carlston and Skowronski (2005) asked participants to familiarize themselves with pairs of behavior descriptions and photos (some were photos of

actors and others of communicators) in a savings-in-relearning paradigm. In the following phase of the experiment, they were shown a photo and a trait and had to rate the extent to which they thought the person in the photo possessed the trait. But before the rating task, the authors asked half of the participants to recall whether the target of the informant's description was the self or the other. They observed that this interposed recall task increased the extremity of ratings made of the actors relative to a control (thus, a higher STI) and reduced the extremity of inferences made about communicators. The STTs were even lower than the control condition, thus totally eliminating them. Hence an effort to recall details of the original descriptions seemed to lead to a double dissociation.

However, this apparent double dissociation with STI and STT is merely an effect on recognition performance, not an effect on the processes at encoding that MATIT is designed to model. Carlston and Skowronski's (2005) manipulation was not originally meant to differentiate STIs and STTs and thus was carried out during the test phase. This is important because the attribution versus associative debate focuses primarily on encoding processes, and no manipulation was conducted to affect encoding in this experiment (see the General Discussion for more on this point). Rating the extent to which the person in the photo possessed the trait is also not the best dependent variable, as it is an explicit task about the formed impression. Our focus and the MATIT model concern how participants make inferences in a spontaneous fashion at encoding.

One way to adapt this manipulation to our goals may be to consider an experiment using the false recognition paradigm, but making participants memorize the photos before they memorize the photo-sentence pairs. Thus, they would first be presented repeatedly with the whole set of photos (one by one), with each as either an actor or an informant about the behavioral descriptions that they will see next. After they memorize who has which role, they would memorize the sentence-photo pairs in the usual way. And finally they would do the recognition test, indicating whether the trait was or was not presented in the sentence. We expect this manipulation to negatively affect the level of false recognitions in STT because the participants would know that it wasn't presented as an actor, creating in this way a biased response toward no-answers ("no the word was not presented with this person"). For STI the same knowledge will work in the opposite direction, knowing that the person in the picture was the actor of the behavior will bias the response toward more false alarms.

Two outcomes seem possible, a double dissociation or a reduced difference between STI and STT. First, better memory of the material may increase the STI effect since participants would know very well that that person actually performed the behavior, whereas STT will decrease since they would know that the implied trait is not related to that person who was only the informant. On the other hand, if the difference between STIs and STTs is simply a matter of the amount of attention paid to the photo, then this procedure should guarantee equivalent attention in both cases, and the difference between STI and STT should disappear. Because we do not have the behavioral evidence from such a study, the question here is whether or not MATIT produces a double dissociation for this particular design.

The simulation is a 2 (Faces presentation: multiple presentations versus standard single presentation) × 2 (Pairing: old versus new pairs) × 2 (Relevance: actor versus non-actor) ANOVA, with the first factor between-Ss and the rest within-Ss.

METHOD

For this simulation, world knowledge is similar to Simulation 1, but there are more input patterns where the photos are repeatedly presented alone (8 times for the actor photo and 8 times for the communicator photo; see Table 8). Thus the frequency for the learning phase remained at 8, and the learning rate was 0.1. The task-specific learning was exactly the same as in Simulation 1, in that actor faces were presented with a higher value (0.5) than non-actor faces (0.25).

SIMULATION RESULTS AND DISCUSSION

We found a main effect of Face Presentation, $F(1, 388) = 42.47$, $p < 0.001$, $\eta^2 = 0.04$; of Pairing, $F(1, 388) = 254.39$, $p < 0.001$, $\eta^2 = 0.19$; and of Relevance, $F(1, 388) = 27.71$, $p < 0.001$, $\eta^2 = 0.02$. The was no significant Presentation × Relevance interaction, $F(1, 388) = 2.15$, $p = 0.15$, $\eta^2 = 0.002$, no significant Presentation × Pairing interaction, $F(1, 388) = 0.02$, $p = 0.89$, $\eta^2 = 0.00$, and also no 3-way interaction, $F(1, 388) = 0.69$, $p = 0.41$, $\eta^2 = 0.00$. The only significant interaction was Relevance × Pairing, $F(1, 388) = 26.91$, $p < 0.001$, $\eta^2 = 0.03$. In terms of the effects we are looking for, as can be seen in Figure 7, and can be concluded from the lack of interctions with the Presentation factor, the main manipulation didn't affect either STI or STT.

There was an STI effect, as in the standard simulation (higher false recognition in old trials than in new ones), $t(49) = 6.74$, $p < 0.001$, $d = 1.38$, 95% CI [0.03, 0.06]; an STT effect, $t(49) = 4.49$, $p < 0.001$, $d = 1.38$, 95% CI [0.03, 0.06]; and a stronger STI than STT effect, $t(49) = 2.31$, $p = 0.03$, $d = 0.50$, 95% CI [0.003, 0.04]. Comparing single (standard simulation) and multiple presentations, the STI effect (old minus new trials) in single presentations was not different from the STI effect in multiple presentations, $t(49) = 0.65$, $p = 0.52$, $d = -0.14$, 95% CI [-0.01, 0.02], and the same was true for the STT, $t(49) = -0.63$, $p = 0.53$, $d = 0.12$, 95% CI [-0.02, 0.01].

These results show that the model cannot produce the elimination of the STT effect and an increase in the STI effect with this multiple photo presentation manipulation. This means that if this double dissociation were obtained experimentally, it would falsify our model. We will turn to this point in the General Discussion.

<A>Simulation 7: Falsification—Double Dissociation 2

This second demonstration has the same aim as the previous one, to examine a plausible double dissociation that, as it turns out, the model cannot simulate. Imagine a hypothetical experiment where we present the same trial twice in the task-specific phase, that is, present each behavior-photo pair twice. This should differently affect STI and STT if attributions are important. The attributional view of STI assumes that the processing of the actor and his/her trait-related behavior

TABLE 8. Learning Pattern: World Knowledge

Behaviors								Traits								Presented person								Non-presented person			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4
0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

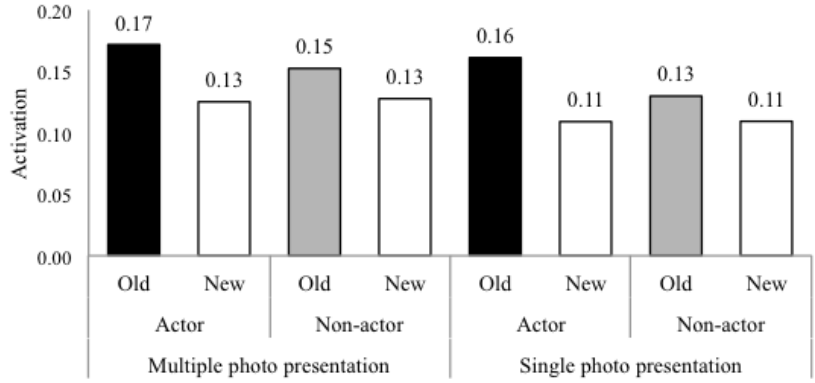

FIGURE 7. Simulation 6: Mean activation of the behavior node.

is deeper and more elaborated than the associative processing of the informant, actor, and behavior in STT. So it is plausible that STI would decrease with an additional exposure to the actor, because participants have the opportunity to better memorize the photo and the behavior and thus to better recall them in the test phase (especially the behavior and the presence/absence of the trait). The better they recall the behavior, the fewer false recognitions they will show. In the STT condition, however, because the processing is more shallow, the actual recollection of the sentence should improve less with the double presentation of the informant photo-behavior pair. The double presentation may nevertheless strengthen the association between the trait and the informant's face, leading to a stronger STT effect.

METHOD

For this simulation, world knowledge is the same as in Simulation 1. The task-specific learning is similar as well. The only difference is the frequency of each pattern of input; two instead of one, because the sentence-photo pairs are presented twice (Table 3). All the rest of the parameters were the same.

SIMULATION RESULTS AND DISCUSSION

The ANOVA on behavior activation showed a main effect of Relevance, $F(1, 388) = 164.60$, $p < 0.001$, $\eta^2 = 0.07$; of Pairing, $F(1, 388) = 1262.97$, $p < 0.001$, $\eta^2 = 0.58$; and no main effect of Repetition, $F(1, 388) = 2.09$, $p = 0.15$, $\eta^2 = 0.00$. The was a significant Repetition × Relevance interaction, $F(1, 388) = 7.88$, $p = 0.01$, $\eta^2 = 0.004$; a significant Repetition × Pairing interaction, $F(1, 388) = 143.94$, $p < 0.001$, $\eta^2 = 0.08$; a significant

Relevance × Pairing interaction, $F(1, 388) = 121.59$, $p < 0.001$, $\eta^2 = 0.08$; and also a three-way interaction, $F(1, 388) = 7.83$, $p = 0.01$, $\eta^2 = 0.01$. As one can see in Figure 8, with repetition of the task-specific learning, there is a strong STI effect, $t(49) = 24.24$, $p < 0.001$, $d = 5.31$, 95% CI [0.10, 0.12]; a strong STT effect, $t(49) = 16.70$, $p < 0.001$, $d = 3.20$, 95% CI [0.05, 0.06]; and a stronger STI than STT effect, $t(49) = 9.48$, $p < 0.001$, $d = 2.00$, 95% CI [0.04, 0.07]. So although the STT increased compared with the standard effect under no repetition, $t(49) = 6.38$, $p < 0.001$, $d = 1.30$, 95% CI [0.02, 0.4], so did the STI effect, $t(49) = 9.51$, $p < 0.001$, $d = 1.98$, 95% CI [0.04, 0.07]. Thus once again the model doesn't produce the decrease in STI and increase in STT as is suggested by the attributional versus associative view.

The first four simulations were developed with an aim of showing what data the model can reproduce (the single dissociations). The two last simulations were conducted to deal with the "overfitting" problem (Roberts & Pashler, 2000), that is, pointing out what data the model would not be able to reproduce (the double dissociations). This is important because every theory or model should provide a way to corroborate or refute itself. Besides illustrating how the model is falsifiable, there is another advantage in thinking about what the model cannot simulate. These two simulations suggest the kind of experiments we should design to seek double dissociations and thus compelling evidence for the existence of two processes or systems (Shallice, 1988). Single dissociations never rule out the possibility of mere quantitative rather than process differences, related in our case to differences in the amount of activation or cognitive resources (e.g., attention, working memory) applied to each task.

GENERAL DISCUSSION

Current theorizing on social inferences explains the findings on spontaneous trait inference (STI) and spontaneous trait transference (STT) with a dual processing approach, that is, attributional and associative processes. The current article presented a computational model (MATIT) framed as a single process approach, the associative process. With this model we sought to demonstrate that the attributional explanation usually offered for the differences between STI and STT is not required by the existing data, and is not even the most parsimonious one. We do not yet have good evidence for dual process claims for STI and STT.

Using simple computer simulation models as gatekeepers that limit premature dual-process conclusions is not new. They have been used in the past to discredit other more complex dual-process models. MINERVA (Hintzmann, 1986; Hintzmann & Ludlum, 1980), that was created to prove that abstract representations weren't necessary to explain the prototype effect, and BEM (Josefowitz, Staddon, & Cerruti, 2009), that was introduced to disprove the need of metacognition in animal behavior, are examples of computer simulations used for this purpose. However, the current article goes beyond this common confirmatory approach and discusses which kind of data would disconfirm MATIT, thereby suggesting experimental designs which would support the dual-process view. We will discuss

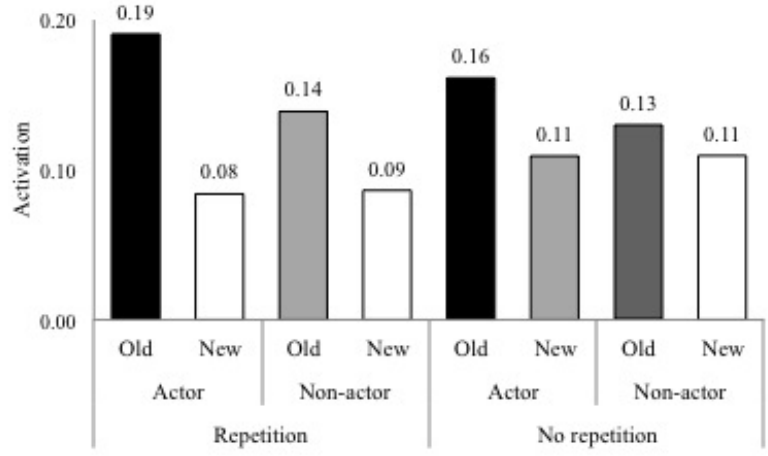

FIGURE 8. Simulation 7: Mean activation of the behavior node

this further at the end of this general discussion. But first we focus the confirmatory part of our work and the theoretical characteristics of MATIT.

Four empirical findings were considered, each of them illustrating a different aspect of the differences between STI and STT. Our hypothesis was that these behavioral data could be simulated by a connectionist model, based on a simple associative learning rule.

The attributional account interprets the difference between STI and STT as being the result of a special linkage which only exists in STI. This link is said to be a "labelled link" created between the actor and the trait, which labels the trait as a property of the actor. In contrast, when the person is not the one who enacts the behavior, a simple association takes place, resulting in a connection only based on space-temporal contiguity. This qualitative difference between STI and STT is said by some authors (e.g., Mae et al., 1999; Skowronski et al., 1998) to account for the larger effects in STI than STT. By contrast, in our associative model, this magnitude difference was ascribed to higher activation values to the actor than to the non-actor in the learning phase. These levels of activation instantiate the presumed levels of attention paid to stimulus persons as a result of the instructions. These in turn influence the weight of the connections between traits and persons, and subsequently produced the difference in magnitude in STI and STT in the first simulation.

Simulation 2 showed that our model can also account for a result where STT is eliminated or reduced. Presenting the actor at the same time as the irrelevant person during encoding the behavioral description eliminates STT and the link between the trait and the non-actor person (Crawford, Skowronski, & Stiff, 2007; Goren & Todorov, 2009; Todorov & Uleman, 2004). This result *per se* is interesting as it shows that this misattribution can be prevented by the presence of the relevant target. In our simulation, the external input consisted of three nodes simultaneously activated, the behavior, the relevant person (with high activation value), and the irrelevant person (with lower activation value). The results of this

implementation went in the same direction as the behavioral data, showing again that it is not necessarily the presence of an attributional process that disrupts the STT effect, but rather differential deployment of node activation and presumably attention.

The third simulation replicates a study where participants are not asked to memorize or familiarize themselves with the material, but rather attempt to judge whether the person in the photo is lying about the behavioral information in the sentence—information that could be about him/herself or about another person. This lie-detection task affects STI but not STT. This has been interpreted as uniquely affecting attributional processes (Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998). But this result was easily simulated by our model. During the encoding phase of the lie-detection condition, we increased the activation values of actor and non-actor to near the same maximum values, to model increased attention to faces in order to detect evidence of lying, and reduced the activation of behaviors to instantiate this presumed shift of attention. This manipulation reduced the activation of the behavior node more for STI than for STT.

The fourth simulation aimed at replicating the halo effect that is more likely to occur in STI than in STT (Carlston & Skowronski, 2005; Crawford, Skowronski, & Stiff, 2007; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski et al., 1998; Wells et al., 2011). This result is said to be attributional because it allows the creation of valence-congruent impressions and generalization to other traits (Carlston & Skowronski, 2005). To simulate the halo effect, we used two types of traits in the world knowledge provided to the model, an implied trait and a valence-consistent trait with associative links with the implied trait. In the test phase, the activation of the behavioral node (the output of the model) was observed in STI and in STT conditions when the consistent trait and the face were presented. Consistent with the experimental data described in the literature, the behavior activation was superior for STI, showing its higher sensitivity to halo effects relative to STT. Crucially, our model explains these findings as a result of the interplay between different associative strengths in world knowledge of how traits relate to behaviors, and task-specific knowledge arising from differential attention to actor and others.

Our simulations were based on variations in the activation values in the learning patterns related to actor and non-actor target. This difference in the activation of actor and non-actor produces differences in the weight of connections (strength of associations) between behaviors and faces, and consequently between faces and traits. Later in the test phase, when we presented only a face and a trait, the activation of the behavioral node depended on connections among nodes and their weights, and these weights usually benefitted STI more than STT.

While we assumed that the (internal) attention paid to the stimuli is the plausible cause of the differences in activation between STI and STT, this is an assumption that needs to be explored. There are two studies (Crawford, Skowronski, Stiff, & Leonards, 2008) that measured visual attention by recording participants' response times to directional probes in various parts of the display (Study 1) or eye movements during encoding (Study 2). Two photos were presented simultaneously with one behavioral description. In some conditions, an *actor* describes his/

her own behavior to a *bystander*; in other conditions, an *informant* describes the behavior of a *target*. The results offered no support for the role of "external attention" in the visual modality (Chun et al., 2011) in producing differences between STI and STT. However this study does not rule out the influence of internal attention.

In another study, Skowronski and collaborators examined whether the STT effect was smaller simply because participants did not pay as much attention to irrelevant behaviors. Behaviors' relevance on each trial in Study 2 was not signaled until participants read it, guaranteeing equal attention under STI and STT conditions. However, the STT effect was not affected by this manipulation, ruling out external attention as a plausible explanation. But their findings do not rule out the influence of internal attention, that is, participants could have stored the displayed information in working memory and only once they knew the relevance of the information was their internal attention directed accordingly. Nevertheless the present article does not provide behavioral support for the internal attention hypothesis. It only suggest it as a plausible cause for the difference in activation of the actor node and the non-actor node. In fact, it is conceivable that other psychological processes (e.g., task setting) can initiate a similar modulation of the activation suggested in our four simulations. For instance, relevance of faces may lead to higher activation of representations whereas irrelevance of faces may lead to lower activation of their representations.

As we stressed through this article, MATIT is a very simple model. For example, MATIT does not address the complex processes by which people construct verbal descriptions of behavior from observations, and infer trait concepts from those behaviors or descriptions. "*Telling the cashier that he received too much change*" is but one way to describe an observed behavior. It might also be described as "*telling the cashier that she had made a mistake*" or "*commenting on the contentiousness of those making minimum wage*" or "*making small talk with the cashier.*" Encoding a behavior into verbal form, and extracting a summarizing trait concept (or gist or goal or style description) involves complex processes and choices (e.g., Semin & Fiedler, 1992). MATIT does not address these.

Finally we return to the theme of falsifying MATIT through exploring plausible data and outcomes that it cannot simulate. In Simulations 6 and 7, we demonstrated that the MATIT model would be falsified by double dissociations. We also argued that a double dissociation is a plausible experimental outcome by drawing on the dual-process attribution versus association account. Thus, one might conclude that a falsification of MATIT by finding a behavioral double dissociation would imply confirmation of the dual-process account. It would not, in part because MATIT can be extended in a sensible way. The autoassociative memory in MATIT has only limited abilities to model complex relationships due to the single layered structure and the linear function governing node activation. This is why it cannot model a double dissociation. Hence, a structure along the lines of a multilayer perceptron/model would be a natural extension, and could model a double dissociation. In fact, multilayer perceptrons are well-known for modelling double dissociations in single process frameworks (e.g., Seidenberg & McClelland, 1989).

But remember that our goal here was not to show that a sophisticated associative model could account for any conceivable data that might be used to support dual-process claims. Our goal was, instead, to show that a very simple associative model that does not posit dual processes could account for major differences between STI and STT that have been interpreted as supporting a dual process account, and to describe some of the consequences of such a model.

REFERENCES

Anderson, C. A. (1995). Implicit theories in broad perspective. *Psychological Inquiry, 6*, 286-290.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249-277.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 834-846.

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

Bassili, J. N. (1976). Temporal and spatial contingencies in the perception of social events. *Journal of Personality and Social Psychology, 33*(6), 680-685.

Bassili, J. N., & Smith, M. C. (1986). On the spontaneity of trait attribution: Converging evidence for the role of cognitive strategy. *Journal of Personality and Social Psychology, 50*(2), 239-245.

Berscheid, E., Graziano, W., Monson, T., & Dermer, M. (1976). Outcome dependency: Attention, attribution, and attraction. *Journal of Personality and Social Psychology, 34*(5), 978-989.

Bodenhausen, G. V., & Hugenberg, K. (2009). Attention, perception, and social cognition. In F. Strack & J. Förster (Eds.), *Social cognition: The basis of human interaction* (pp. 1-22). Philadelphia: Psychology Press.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214-234.

Brown, R. D., & Bassili, J. N. (2002). Spontaneous trait associations and the case of the superstitious banana. *Journal of Experimental Social Psychology, 38*(1), 87-92.

Bundesen, C., Habekost, T., & Kylingsbaek, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review, 112*(2), 291-328.

Carlston, D. E., & Skowronski, J. J. (2005). Linking versus thinking: Evidence for the different associative and attributional bases of spontaneous trait transference and spontaneous trait inference. *Journal of Personality and Social Psychology, 89*(6), 884-898.

Carlston, D. E., Skowronski, J. J., & Sparks, C. (1995). Savings in relearning: II. On the formation of behaviour-based trait associations and inferences. *Journal of Personality and Social Psychology, 69*(3), 420-436.

Carlston, D. E., & Smith, E. R. (1996). Principles of mental representation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 184-210). New York: Guilford.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature, 363*, 345-347.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73-101.

Clary, E. G., & Tesser, A. (1983). Reactions to unexpected events. *Personality and Social Psychology Bulletin, 9*(4), 609-620.

Craik, F. I. M., & Lockhart, R. S., (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11*, 671-684.

Crawford, M. T., Skowronski, J. J., & Stiff, C. (2007). Limiting the spread of spontaneous trait transference. *Journal of Experimental Social Psychology, 43*(3), 466-472.

Crawford, M. T., Skowronski, J. J., Stiff, C., & Leonards, U. (2008). Seeing, but not thinking: Limiting the spread of spontaneous trait transference II. *Journal of Experimental Social Psychology, 44*(3), 840-847.

Crawford, M. T., Skowronski, J. J., Stiff, C., & Scherer, C. R. (2007). Interfering with inferential, but not associative, processes underlying spontaneous trait inference. *Personality and Social Psychology Bulletin, 33*(5), 677-690.

Davies, M. (2010) Double dissociation: Understanding its role in cognitive neuropsychology. *Mind and Language, 25*(5), 500-540.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology, 36*(5), 451-462.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39*, 1-7.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behaviour. *Journal of Personality and Social Psychology, 38*, 889-906.

Fiske, S., Kenny, D. A., & Taylor, S. E. (1982). Structural models for the mediation of salience effects. *Journal of Experimental Social Psychology, 18*, 105-127.

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Fox, E., Derakshan, N., & Standage, H. (2011). The assessment of human attention. In K. C. Klauer, A. Voss, & C. Stahl (Eds.), *Cognitive methods in social psychology* (pp. 15-47). New York: Guilford.

Garcia-Marques, L., & Ferreira, M. B. (2011). Friends and foes of theory construction in psychological science: Vague dichotomies, unified theories of cognition, and the new experimentalism. *Prespectives on Psychological Science, 6*(2), 192-201.

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

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*(5), 733-740.

Goren, A., & Todorov, A. (2009). Two faces are better than one: Eliminating false trait associations with faces. *Social Cognition, 27*(2), 222-248.

Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is "the fundamental attribution error"? *Journal of Personality and Social Psychology, 40*(2), 346.

Hassin, R. R., Aarts, H., & Fergunson, M. L. (2005). Automatic goal inferences. *Journal of Experimental Social Psychology, 41*, 129-140.

Hassin, R., Bargh, J. A., & Uleman, J. S. (2002). Spontaneous causal inferences. *Journal of Experimental Social Psychology, 38*, 515-522.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology, 46*(1), 44-56.

Heider, R. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Heider, F. (1982). *The psychology of interpersonal relations*: Hillsdale, NJ: Erlbaum.

Heinke, D. (2009). Computational modelling in behavioural neuroscience: Methodologies and approaches—Minutes of discussions at the workshop in Birmingham, UK in May 2007. In D. Heinke & E. Mavritsaki (Eds.), *Computational modelling in behavioural neuroscience: Closing the gap between neurophysiology and behaviour* (pp. 332-338). London: Psychology Press.

Heinke, D., & Backhaus, A. (2011). Modeling visual search with the Selective Attention for Identification model (VS-SAIM): A novel explanation for visual search asymmetries. *Cognitive Computation, 3*(1), 185-205.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological review, 93*(4), 411-428.

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory and Cognition, 8*, 378-382.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology, 2*, 219-266.

Jozefowiez, J., Staddon, J. E. R., & Cerutti, D. T. (2009). Metacognition in animals: How

do we know that they know? *Comparative Brain and Behaviour Reviews, 4*, 19-29.

Kanazawa, S. (1992). Outcome or expectancy? Antecedent of spontaneous causal attribution. *Personality and Social Psychology Bulletin, 18*(6), 659-668.

Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review, 107*(4), 914-942.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron, 22*, 751-761.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15*, 192-238.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107-128.

Kressel, L. M. (2011). *The functional meaning of traits and spontaneous trait inferences.* Unpublished dissertation, New York University.

Kressel, L., & Uleman, J. S. (2010). Personality traits function as causal concepts. *Journal of Experimental Social Psychology, 46*, 213-216.

Lau, R. R., & Russell, D. (1980). Attributions in the sports pages. *Journal of Personality and Social Psychology, 39*(1), 29-38.

Lepsien, J., & Nobre, A. C. (2006). Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. *Brain Research, 1105*, 20-31.

Maass, A., Colombo, A., Colombo, A., & Sherman, S. J. (2001). Inferring traits from behaviors versus behaviors from traits: The induction-deduction asymmetry. *Journal of Personality and Social Psychology, 81*(3), 391-404

Mae, L., Carlston, D. E., & Skowronski, J. J. (1999). Spontaneous trait transference to familiar communications: Is a little knowledge a dangerous thing? *Journal of Personality and Social Psychology, 77*(2), 233-246.

Malle, B. F. (2003). *Attributions as behavior explanations: Toward a new theory.* Unpublished manuscript, University of Oregon, Eugene, OR.

Mavritsaki, E., Heinke, D., Allen, H., Deco, G., & Humphreys, G. W. (2011). Bridging the gap between physiology and behavior: Evidence from the sSoTS model of human visual attention. *Psychological Review, 118*(1), 3-41.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159-188.

McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing-Macintosh version: A handbook of models, programs, and exercises.* Cambridge, MA: MIT Press.

Naveh-Benjamin, M., Craik, F.I.M., Perratta, J., & Tonev, S.T. (2000). The effects of divided attention on encoding and retrieval processes: The resiliency of retrieval processes. *Quarterly Journal of Experimental Psychology, 53*, 609-626.

Orghian, D., Gancarcyk, S., Garcia-Marques, L., & Heinke, D. (2014). *Why "well done is better than well said": The involvement of attention in spontaneous trait inferences and spontaneous trait transferences.* Manuscript submitted for publication.

Pashler, H. (1995). Divided visual attention. In S. Kosslyn (Ed.), *Visual cognition: Invitation to cognitive science* (pp. 71-100). Cambridge, MA: MIT Press.

Pittman, T. S., & Pittman, N. L. (1980). Deprivation of control and the attribution process. *Journal of Personality and Social Psychology, 39*(3), 377-389.

Pyszczynski, T. A., & Greenberg, J. (1981). Role of disconfirmed expectancies in the instigation of attributional processing. *Journal of Personality and Social Psychology, 40*(1), 31-38.

Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's (1998) critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology, 76*(5), 728-742.

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory

testing. *Psychological Review, 107*(2), 358-367.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 501-518.

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79*, 294-309.

Schneider, D. J. (2004). *The psychology of stereotyping*. New York: Guilford.

Seidenberg, M., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523-568.

Semin, G. R., & Fiedler, K. (Eds.). (1992). *Language, interaction and social cognition*. London: Sage.

Shallice, T. (1988). *From neuropsychology to mental srucutre.* Cambridge: Cambridge University Press.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*(2), 219-240.

Singer, M. (1979). Process of inferences in sentence encoding. *Memory and Cognition, 7*, 192-200.

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: Communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology, 74*(4), 837-848.

Smith E. R., & DeCoster, J. (1999). Associative and rule-based processing: A connectionist interpretation of dual process models. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 323-336). New York: Guilford.

Smith, E. R., & Miller, F. D. (1983). Mediation among attributional inferences and comprehension processes: Initial findings and a general method. *Journal of Personality and Social Psychology, 44*(3), 492-505.

Swann, W. B., Stephenson, B., & Pittman, T. S. (1981). Curiosity and control: On the determinants of the search for social knowledge. *Journal of Personality and Social Psychology, 40*(4), 635-642.

Taylor, S. E., & Fiske, S. T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology, 32,* 439-445.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 11* (pp. 249-288). New York: Academic Press.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051-1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549-562.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87*(4), 482-493.

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. *Advances in Experimental Social Psychology, 28*, 211-279.

Uleman, J. S., Rim, S., Saribay, S. A., & Kressel, L. M. (2012). Controversies, questions, and prospects for spontaneous social inferences. *Social and Personality Psychology Compass, 6*, 657-673.

Van Overwalle, F. (1998). Causal explanation as constraint satisfaction: A critique and a feedforward connectionist alternative. *Journal of Personality and Social Psychology, 74*(2), 312-328.

Van Overwalle, F., & Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review, 6*(3), 204-231.

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review, 8*(1), 28-61.

Wells, B. M., Skowronski, J. J., Crawford, M. T., Scherer, C. R., & Carlston, D. E. (2011). Inference making and linking both require thinking: Spontaneous trait inference and spontaneous trait transference both rely on working memory capacity. *Journal of Experimental Social Psychology, 47*, 1116-1126.

Wexler, K. (1978). A review of John R. Anderson's language, memory, and thought. *Cognition, 6*, 327-351.

Wexler, P. (1987). *Social analysis of education: After the new sociology.* London: Routledge & Kegan Paul.

Wigboldus, D. H. J., Dijksterhuis, A., & Knippenberg, A. V. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology, 84*(3), 470-484.

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

Wong, P. T., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology, 40*(4), 650-663.

Yoon, E. Y., Heinke, D., & Humphreys, D. (2002). Modelling direct perceptual constraints on action selection: The Naming and Action Model (NAM). *Visual Cognition, 9*(4-5), 615-661.

APPENDIX A. DETAILS OF THE AUTOASSOCIATIVE MODEL

Each node in the model represents a construct with psychologically interpretable meaning. The activation of these nodes leaves a "memory trace" behind that results from changes in the weights of the connections between nodes, that is, changes in the strength of these connections that are responsible for the learning gains of the model (McClelland & Rumelhart, 1985). These weighted connections between units store the information required to complete familiar (learned) patterns.

During the recall phase, the network receives activation from the exterior—external input. But nodes also receive input from the other nodes in the network—internal input. Considering two nodes, i and j, the input from j to i is i_{ij} and is:

$i_{ij} = a_j w_{ij}$

where a_j is the activation of the node j, and w_{ij} is the weight that defines the influence of node j on node i.

The internal input is the sum of the activation coming from all the other nodes on the network:

$int_i = \sum(a_j w_{ij})$.

The sum of the external and internal inputs is the net input of the node:

$net_i = ext_i + int_i$

where ext_i is the external input and int_i is the internal input.

The memory trace is created so as to better anticipate and characterize the future external input. This memory trace is constructed from the discrepancy between the internal input of the network from the last updating cycle and the external input. Mathematically, weights between nodes are adjusted by the delta rule algorithm (McClelland & Rumelhart, 1989):

$\Delta w_{ij} = \varepsilon(ext_i - int_i) a_j$

where w_{ij} is the connection's weight from j to i, e is the learning rate that defines the learning speed of the model, and a_j is the activation of the node j. In a model with enough learning trials and reasonable learning rate, the Δw_{ij} will tend to zero as the model reaches a stable state. In such a state, the model can anticipate efficiently the input received from the external environment.

Thus, the learning in the model is accomplished through the computation of errors and the updating of the weights between nodes so as to minimize this error.

A linear version of the autoassociative network was applied the current work, and differently from McClelland & Rumelhart (1985), we used only one internal updating cycle rule proposed by Van Overwalle and Labiouse (2004, p. 60) which allows for faster and simpler simulations.

For those unfamiliar with such models, a walk-through example may be helpful, keyed to the false recognition paradigm of Todorov and Uleman (2002, 2003, 2004). Subjects view a series of stimuli, each containing a photo of a person's face and a sentence (on critical trials) describing a trait-implying behavior. They view these stimuli in this first phase in preparation for "a memory test" in the second phase of the paradigm. Some of the sentences contain the trait explicitly, but these merely set up the subsequent false recognition test in which subjects have to remember whether or not the trait was explicitly in the sentence. In the second phase, each memory test item re-presents a person photo paired with a trait term, and subjects must judge whether the trait was explicitly in the sentence they saw earlier. On critical trials, the trait was merely implicit so the correct answer is "No." False recognition errors ("Yes," with appropriate controls) measure the extent to which subjects spontaneously (i.e., unintentionally and unconsciously) inferred traits during the first phase.

Therefore in the model above, there are three concepts (or open circle nodes): B, the trait-implying behavior; T, the trait implied; and F, the person's face. They can each receive external input, and are connected through bi-directional links to each other. So every node is potentially connected to every other node as well as to the input and output signals. The connections are "potential" because links can vary in conductivity from 0 (not connected) to 1 (connected with complete conductivity or no resistance). A "weight" describes the conductivity of each link, and varies with each trial according to the delta learning algorithm. The algorithm adjusts the weights in the network so that the resultant activation of the nodes on trial $n + 1$ matches more closely their activation on trial n. Activation is introduced into the system through the "external input."

For simplicity's sake, we considered only the effects of presenting photos with trait-implying (not trait-explicit) behaviors. Links among nodes in the model are initially set to zero, so that no activation is transmitted from node to node. Trait inference depends on "teaching" the model the world knowledge that particular behaviors are associated with particular traits, and that traits can therefore be "inferred" from these behaviors. Subjects enter the study with this world knowledge, but it must be imparted to the model via the input matrix in Table 2. This shows 8 behaviors, 8 traits, and 8 faces in the model (so the actual running model is more complex than the simple figure). The model "learns" slowly and imperfectly, so that links among nodes are never completely conductive with weights of 1. In Simulation 1 (above), world knowledge was imparted by presenting the matrices of Table 2 as input 8 times. In the first presentation, behaviors 1 through 8 are activated by external input, each along with its corresponding trait. (Faces were not activated because world knowledge is about behavior-trait implications, not knowledge about who did what). Imagining that the initial weight between nodes is zero, in the first trial, the delta learning algorithm adjusted the zero weights between pairs of behaviors and traits slightly from zero to a value of .0025. (The change in weight, Δw, is given by ε [ext – int] a, where ε = .01, ext = 0.5, int = 0, and a = 0.5. Note that although the links are potentially asymmetric, in that activation from node i to j need not be the same as activation from j to i, we've dropped this feature here for simplicity and because it does not affect these simulations.)

After this first input of world knowledge, nodes receive both external activation (ext) from a repetition of Table 2's paired behaviors and traits (7 more times), and internal activation (int) from other nodes; so calculating Δw becomes slightly more complex. After the second input, $\Delta w = .01 \times (0.5 - .0025) \times 0.5 = .00248$, raising w between behavior and trait to .0050. Δw decreases with each new input, so that after 8 inputs, $w \approx .020$, the conductivity of the link connecting pairs of behaviors and traits. Thus the pairs of behaviors and traits become linked in the model through their simultaneous activation and the delta learning rule, which moves the model's internal activation values toward the external activation values that it has experienced.

Now that the model "knows" as much as the subjects do, it can participate in the first (study) phase of the false recognition paradigm, learning behavior-face pairs. This occurs through input matrices like Table 2 but with different values, as in Table 3 for Simulation 1.

Note that all the values for the traits in the matrix are zero, because traits are never presented, only face-behavior pairs. This produces some activation of trait nodes because world knowledge links behaviors and traits. It also associates faces and traits because faces and trait-implying behaviors occur simultaneously. The input matrices are presented once, just as the stimuli are presented to subjects once. Then the test phase follows, in which subjects are asked whether particular traits appeared in behavioral sentences with particular photos. This test is simulated by presenting the model with face-trait pairs and seeing how much they activate the corresponding behavior nodes (because the question is whether or not the trait appeared in the behavior). That is, behavior node activation is read off, and serves as the dependent variable in the false recognition paradigm.