Social Cognition, Vol. 38, No. 2, 2020, pp. 146-178

BEING HONEST AND ACTING CONSISTENTLY: BOUNDARY CONDITIONS OF THE NEGATIVITY EFFECT IN THE ATTRIBUTION OF MORALITY

Patrice Rusconi
University of Surrey

Simona Sacchi, Marco Brambilla, Roberta Capellini, and Paolo Cherubini
University of Milano-Bicocca

Morality, which refers to characteristics such as trustworthiness and honesty, has a primary role in social perception and judgment. A negativity effect characterizes the morality dimension, whereby negative information is weighed more than positive information in trait attribution and impression formation. This article reviews the literature on the negativity effect in trait attribution and impression formation. We examine the main boundary conditions of the negativity effect by considering relevant moderators, such as behavior consistency and evaluative extremity, level of categorization, and measurement type as well as some theoretical and empirical inconsistencies in the literature. We also review recent studies showing that social perceivers hold negative assumptions about people's morality. We outline future directions for research on the negativity effect that should consider trait extremity, use alternative measures to the perceived frequency of behaviors, introduce more precise definitions of relevant constructs, such as diagnosticity, and test different schemata of trait-behavior relations.

Keywords: negativity effect, morality, trait attribution, impression formation

All the authors conceived the research idea. P. Rusconi conducted the literature review and drafted the first version of the article, while the other authors read and commented on it.

We would like to thank John Skowronski and an anonymous reviewer for insightful comments and suggestions on the article.

Correspondence concerning this article should be addressed to Patrice Rusconi, School of Psychology, Department of Psychological Sciences, Faculty of Health and Medical Sciences, Room 15AC05, University of Surrey, Guildford, Surrey, UK, GU2 7XH. E-mail: p.rusconi@surrey.ac.uk

Studies on morality in trait attribution, social perception, and judgment have burgeoned in the last decade (e.g., Brambilla & Leach, 2014; Ellemers, van der Toorn, Paunov, & van Leeuwen, 2019; Goodwin, 2015; Leach, Ellemers, & Barreto, 2007; Mende-Siedlecki, Baron, & Todorov, 2013; Trafimow, Bromgard, Finlay, & Ketelaar, 2005). Research in this field has shown that characteristics related to morality such as honesty and trustworthiness, which refer to human benevolence and correctness in social interactions, have a primary role in shaping the perceptions of ourselves and others (Brambilla & Leach, 2014; Brambilla, Rusconi, Sacci, & Cherubini, 2011; Goodwin, Piazza, & Rozin, 2014; Landy, Piazza, & Goodwin, 2016, 2018; Leach et al., 2007). This primacy of morality has been interpreted from a sociofunctionalist perspective as originating from a motivation to establish whether someone represents an opportunity or a threat for the self (Brambilla, Biella, & Freeman, 2018; Brambilla et al., 2011; Brambilla, Sacchi, Rusconi, Cherubini, & Yzerbyt, 2012; Brambilla, Sacchi, Pagliaro, & Ellemers, 2013; De Bruin & Van Lange, 2000; Rusconi, Sacchi, Capellini, Brambilla, & Cherubini, 2017; Todorov, Baron, & Oosterhof, 2008; Todorov, Said, Engell, & Oosterhof, 2008). Further, an asymmetry characterizes the positive and negative poles of the morality dimension. Indeed, a negativity effect has been found, whereby negative information is weighed more than positive information when making trait judgments and forming impressions of other people's morality (e.g., Peeters & Czapiński, 1990; Skowronski & Carlston, 1989). The literature has pointed out several mechanisms that might contribute to the explanation of the negativity effect (e.g., Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Taylor, 1991) revealing the complex nature of this asymmetry (Peeters & Czapiński, 1990).

Previous reviews on the negativity effect were broad in scope as they considered a wide range of phenomena and processes (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Peeters & Czapiński, 1990; Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Taylor, 1991). Building on these prior works, our review aims to provide an updated description of the different models on the negativity effect, showing that these different accounts are not necessarily in contradiction to one another. Moreover, we intend to highlight some empirical inconsistencies about the impact of negative moral information on social judgment. We argue that research on the negativity effect, especially in the morality domain, should consider relevant boundary conditions (cognitive and affective moderators) that could account for those contradictions. By doing so, we will highlight some open issues and suggest future research avenues that might contribute to extend this area of social cognition.

THE NEGATIVITY EFFECT AND ITS ACCOUNTS

An information-processing bias affects information integration, impression formation, and evaluative processes in general, whereby social perceivers assign more weight to negative information and events than positive ones of equal intensity. For example, gaining enemies would be weighed more than gaining friends (Baumeister et al., 2001; Kanouse, 1984; Kanouse & Hanson, 1972; Peeters & Czapiński,

1990; Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Taylor, 1991; Ybarra, 2002).

This negativity effect challenges the basic version of an averaging model of impression formation that assumes that equal weights are assigned to the adjectives on which an impression is based (Anderson, 1965, 1968; Anderson & Alexander, 1971). Anderson already noted that findings departing from the simplified version of the averaging model could be explained by people's tendency to assign more weight to extremely negative information (Anderson & Alexander, 1971, p. 314).

Several explanations have been put forward to explain the enhanced weight to negative information (for previous reviews, see Baumeister et al., 2001; Kanouse, 1984; Peeters & Czapiński, 1990; Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Taylor, 1991; see also Trafimow et al., 2005). Kanouse (1984) classified them in "micro-level" and "macro-level" accounts. Micro-level accounts focus on processes: perception, attention, memory, and judgment. Macro-level accounts focus on the social perceiver's goals and the wider context. In addition to these accounts, the negativity effect has been explained with reference to cognitive processes and variables (e.g., constructs, previous knowledge, diagnosticity and implicit assumptions; see, for example, Reeder & Brewer, 1979; Reeder, Henderson, & Sullivan, 1982; Skowronski & Carlston, 1987, 1989), and socio-functionalist and affective processes (e.g., Ito, Larsen, Smith, & Cacioppo, 1998; Peeters & Czapiński, 1990; Reeder, 2009; Trafimow et al., 2005). In the present review, we will categorize these accounts into cognitive explanations, encompassing both cognitive processes and cognitive structures, and motivational and affective explanations, based on evaluative (e.g., like-dislike) judgments and the social perceiver's reactions (e.g., arousal). We will also consider the interplay between cognitive and motivational accounts.

COGNITIVE EXPLANATIONS

Novelty. Within the cognitive explanations of the negativity effect, some theories focused on the role of social perceiver's expectations and the contrast between those a priori beliefs and the social stimuli. Some of these accounts rely on perceptual contrast and, in particular, on a comparison between objective stimuli and the perceiver's anchors (Helson, 1947, 1948; Sherif & Sherif, 1967). Negative stimuli contrast more strongly than positive stimuli do given that social perceivers are anchored on moderately positive expectations. For example, a dishonest behavior would be evaluated more negatively than it objectively should be because it is distant from the subjective anchor which is shifted toward a moderately positive (moral) behavior.

In a similar vein, Fiske (1980) emphasized the role of novel information by focusing on attentional processes. She found that attention (operationalized as looking times) taps into the differential weight participants assigned to negative, as opposed to positive, information. Fiske's (1980) novelty approach assumes that social perceivers generally hold moderately positive expectations about other

people's behaviors. Thus, a negative behavior would be novel, unexpected, and consequently more informative and influential on impressions. The same assumption based on the discrepancy between the social perceiver's expectations and other people's behaviors is shared by Jones and Davis's (1965) correspondent inference theory. Jones and co-authors (Jones, 1976; Jones & Harris, 1967) show that during the impression-formation process on social targets the social perceiver assigns a greater weight to information and behaviors that do not conform to social norms, thus being unexpected, distinct and informative (Jones & McGillis, 1976; see also Skowronski & Carlston, 1989).

Fiske's (1980) and Jones and colleagues' (Jones & Davis, 1965; Jones & McGillis, 1976) accounts both rely on violations of expectations in terms of behaviors' novelty or non-normativity. They can thus be classified under the same umbrella (Skowronski & Carlston, 1989 labeled them "frequency-weight" theories). However, these two theories differ because Fiske's (1980) approach focuses on the role of attention, whereas Jones and colleagues' (Jones & Davis, 1965; Jones & McGillis, 1976) theory is based on the social perceiver's constructs about the social desirability of behaviors. Both accounts have been criticized for the lack of direct evidence for the role of novelty and non-normativity in causing the negativity effect (Skowronski & Carlston, 1989). As such, Pratto and John (1991), in a color-naming task, showed participants' greater latency in naming the color of undesirable trait adjectives (e.g., "mean") than of positive traits (e.g., "sincere"). This finding was interpreted as evidence for an automatic attentional processing of negative stimuli. The authors also found an incidental learning effect, whereby participants, on average, recalled twice as many undesirable traits as desirable traits (Pratto & John, 1991, Experiment 2). This negativity effect in attention and memory was not explained by the relative greater novelty of undesirable traits compared to desirable traits because the manipulation of perceived trait base rates did not yield any significant effects (Pratto & John, 1991, Experiment 3).

Range Underlying Behaviors and Diagnosticity. A second class of cognitive explanations for the negativity effect are the theories that focus on the uncertainty underlying behaviors and how people gauge it to form an impression (Birnbaum, 1972; Wyer, 1974). These "range" theories assume that social behaviors are underlain by a range of possible values, and the width of this range of values determines the behavior's level of ambiguity. The narrower the distribution of values underlying the judgment of a behavior, the less ambiguous and more influential that behavior is. Negative stimuli are characterized by narrower ranges because they are perceived as less various and ambiguous, thus being weighed more in people's impressions than positive stimuli (Birnbaum, 1972, 1974; Birnbaum, Parducci, & Gifford, 1971; Wyer, 1974). In a comparison of averaging and range models, Birnbaum (1972) found that the latter better captured ratings ("how wrong that *pair* of actions would be in terms of your own personal set of values," Birnbaum, 1972, p. 36) of pairs of immoral behaviors (e.g., "poisoning a neighbor's dog whose barking bothers you," "keeping a dime you find in a telephone booth," Birnbaum, 1972, p. 36) than averaging models did. Birnbaum advanced an "overlap of value

hypothesis" to interpret these results. According to this hypothesis, perceivers' pair ratings would be based on the range and mean of the overlapping values between the two distributions of possible values underlying those behaviors. The narrowest of the two distributions of possible values (the one of the more immoral behavior) would thus shift the final value of the pair rating. A limitation of the range theories is that they do not provide an explanatory mechanism for the reduced uncertainty that characterizes negative information (Skowronski & Carlston, 1989).

The cognitive explanations analyzed above share a criticism: They do not take into account the moderating role of trait content (Skowronski & Carlston, 1987, 1989; Trafimow et al., 2005). In contrast, two complementary theories can account for the differences between, for example, competence-related and morality-related traits: The schematic model of trait attribution (Reeder & Brewer, 1979; Reeder et al., 1982) and the cue-diagnosticity account (Skowronski & Carlston, 1987).

The Schematic Model. Reeder and colleagues analyzed how traits related to morality, competence, and preference differ in relation to behavioral expectations (also called "implicational schemata"; Reeder, 1993, 1997, 2006; Reeder, Pryor, & Wojciszke, 1992; Reeder & Brewer, 1979; Reeder et al., 1982). Trait-behavior relations are abstract assumptions that people hold on the range of behaviors associated with a trait. For example, one of these implicational schemata is the "hierarchically restrictive schema" (Reeder & Brewer, 1979; Reeder et al., 1982), whereby social perceivers assume that an honest person would attempt to refrain from any dishonest behaviors, thus she/he would be restricted to honest behaviors only. Conversely, a dishonest person is thought to try to behave across the whole spectrum, from dishonest to honest behaviors, thus she/he is behaviorally unrestricted (see Figure 1). A reverse asymmetry characterizes the competence dimension. For example, intelligent people are thought capable of both intelligent and unintelligent behaviors depending on the circumstances, but unintelligent people are not thought capable of intelligent behaviors (Reeder et al., 1982, p. 357). In other words, echoing Woody Allen: "The advantage of being intelligent is that we can always play stupid; however, the opposite is completely impossible." These trait-behavior relations play an important role in trait attribution since the trait inference based on behaviors related to the unrestrictive pole (e.g., immorality- and competence-related behaviors) is enhanced (Reeder, 1993, 1997, 2006).

The Cue-Diagnosticity Account. The cue-diagnosticity approach (Skowronski & Carlston, 1987, see also 1989) builds on the schematic model by Reeder and colleagues, and it focuses on how a cue (i.e., a behavior) can distinguish between alternative categorizations (i.e., the actor has or does not have a specific trait; see, for example, Skowronski, 2002). This categorization is influenced by the diagnosticity of the behaviors or cues used to make the decision. Cue-diagnosticity refers to the ability of a cue to discriminate between two alternative trait categories: The more frequently and exclusively a cue is linked to the members of a target category compared to its alternative, the more diagnostic it is (e.g., Skowronski, 2002).

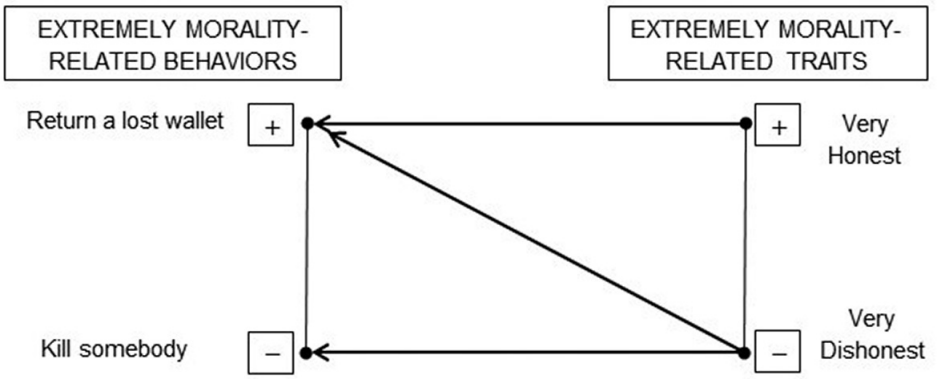

FIGURE 1. Representation of the hierarchical restrictive schema for extreme levels of positive and negative morality-related traits and behaviors according to Reeder and Brewer's (1979) model. + and - indicate traits/behaviors of extremely positive and negative valence, respectively (moderate levels are not included in this representation). The solid arrows indicate strong trait-behavior relations.

Skowronski and Carlston (1987) used a normalized proportion encompassing the perceived likelihoods of trait-inconsistent and trait-consistent behaviors as a measure of the diagnosticity of a behavior for a trait category (the "cue-validity index"). They found that people perceived negative morality behaviors and positive competence behaviors as more diagnostic than their opposites (Skowronski & Carlston, 1987, Experiment 1). Further, perceived behavior diagnosticity, as measured by cue validity, mimicked the pattern of participants' perceived likelihoods of trait-inconsistent behaviors (Skowronski & Carlston, 1987, Figures 1–2). This similarity between the results yielded by the cue-validity index and the likelihood of trait-inconsistent behaviors shows the tight connection between the cue-diagnosticity account and Reeder and Brewer's (1979) schematic model, which also describes the same positivity and negativity effects in relation to trait-inconsistent behaviors along the competence and morality dimensions (e.g., Singh & Teoh, 2000).

The positivity effect in the competence domain, whereby positive information about a person's competence (e.g., a high performance on a difficult test) carries more weight than negative information (e.g., a failure) on impressions (Heider, 1958; Reeder et al., 1982; Skowronski & Carlston, 1987, 1989), is problematic in the perspective of the novelty approach (Fiske, 1980), the correspondent inference theory (Jones & Davis, 1965; Jones & McGillis, 1976), and the range theories (Birnbaum, 1972; Wyer, 1974) that cannot account for this content asymmetry. Indeed, these theories focus on the role played by negative stimuli, in terms of their contrast to the social perceiver's moderately positive expectations (Helson, 1947, 1948; Sherif & Sherif, 1967), their novelty (Fiske, 1980), non-normativity (Jones & Davis, 1965), and reduced ambiguity (Birnbaum, 1972; Wyer, 1974), thus being unable to account for the situations in which positive information has an enhanced weight

(e.g., Trafimow et al., 2005). However, the positivity effect in the competence domain can be accounted for by the schematic model and the cue-diagnosticity account.

An alternative to these accounts that are based on the informativeness of behaviors is represented by the frequency-based theories.

Frequency. The studies by Mende-Siedlecki and co-authors (2013) and Sanbonmatsu and colleagues (2015) indicate that behavior frequency and perceived rarity might contribute to explaining both negativity and positivity effects in impression formation (Mende-Siedlecki et al., 2013; Sanbonmatsu, Mazur, Behrends, & Moore, 2015). For example, in a social neuroscience research, Mende-Siedlecki and co-authors tested the hypothesis that the perceived frequency of behaviors determines participants' updates of initial impressions after receiving contradictory behavioral evidence (Mende-Siedlecki et al., 2013). They found that positive behaviors in the competence domain and negative behaviors in the morality domain were perceived as less frequent than the corresponding behaviors of the opposite valence. Further, these less frequent behaviors determined larger absolute changes in impression ratings about a target person than their correspondent, opposite behaviors did.

In a similar way, Sanbonmatsu and colleagues (2015) focused on the perceived base-rate frequency of trait-consistent behaviors (e.g., how frequently people behave aggressively when given the chance to behave either aggressively or not aggressively) as the process at the basis of trait attribution and the differential attributional weight assigned to immoral versus moral and competent versus incompetent behaviors. They found that the perceived commonality or rarity of behaviors determined the number of instances needed to confirm or disconfirm a trait via the mediation of participants' behavioral expectations (i.e., the social perceiver's assumptions about the behaviors expected from an actor with a trait).

These works are not in contradiction with Fiske's (1980) and Jones and colleagues' (Jones & Davis, 1965; Jones & McGillis, 1976) explanations, since the violation of normative behaviors not only contravenes expectations but is also rarer than normative-consistent behaviors. Thus, all these works build on the notion of perceived rarity, more precisely on Kelley's (1973) criterion of consensus (how common a response to a stimulus is).

MOTIVATIONAL AND AFFECTIVE EXPLANATIONS

We review the motivational and affective explanations of the negativity effect together because some theories interpret the social perceivers' affective reactions as intertwined with their motivation to adaptively respond to negative events (see the approach-avoidance continuum of behaviors in Peeters & Czapiński, 1990; for a similar categorization of affect-based explanations, see Skowronski, 2002). The motivational and affective explanations of the negativity effect consider both the micro level of physiological responses and the macro level of the social perceivers' motivations to avoid threatening stimuli, as well as their affective reactions to

other people's negative behaviors. These views are compatible with the cognitive ones although the underlying mechanism that is hypothesized is different.

Neuro-Physiological Accounts. The role of motivation has been highlighted by Cacioppo and colleagues who proposed specific motivational systems underlying positivity and negativity effects (Cacioppo & Berntson, 1994; Cacioppo, Gardner, & Berntson, 1997; Ito, Larsen et al., 1998). The authors criticized the *bivariate evaluative plane*, whereby the positive and negative evaluative processes fall along a single, bipolar continuum (for example, from "hostile" to "hospitable," Cacioppo et al., 1997) and they are reciprocally activated because they are the endpoints (i.e., very positive and very negative) of the same continuum (Cacioppo & Berntson, 1994). In this sense, positive and negative evaluative processes underlying attitudes are also interchangeable because when one of the two increases, the other one decreases. For example, if a stimulus is regarded as maximally hostile, it will also be evaluated as minimally hospitable. Instead of this single, bipolar conceptualization of the evaluative processes underlying attitudes, the authors introduced a *bivariate model of evaluative space*, whereby positive and negative evaluative processes are underlain by relatively separable neural correlates and motivational systems that guide the social perceiver's action toward a goal. In addition to the reciprocal (bipolar) activation of positive and negative evaluative processes, this model of evaluative space allows for other combinations of positive and negative activations (i.e., negative, positive, and non-significant correlations; Ito, Cacioppo, & Lang, 1998). This model explains the negativity effect in terms of different activation functions underlying positivity and negativity, whereby the activating input increases the motivational output to a higher rate for negativity than for positivity (Ito, Cacioppo et al., 1998). In other words, the negativity effect is due to the greater response of the negative as opposed to the positive motivational system to equivalent amounts of activation (Ito, Larsen, et al., 1998). Without any additional specifications, the model of evaluative space does not account for the moderating role of trait content (e.g., morality vs. competence).

Taylor (1991) has also discussed the greater physiological arousal associated with negatively valenced stimuli in terms of a short-term mobilization that animals and humans undergo when they initially respond to negative events. After this initial phase, an automatic physiological response triggers the minimization of the negative events. The minimization of the negativity effect can also occur due to motivational factors, as reviewed by Baumeister and colleagues (2001). An example is the fading affect bias, whereby the emotional intensity for autobiographical memories of unpleasant events fades faster than that for pleasant memories over time (Walker, Vogl, & Thompson, 1997). The authors interpreted this finding in terms of people's motivation to minimize the negative affect associated with one's memory to protect the self and to manage others' impression by presenting a positive self-image.

In line with this interpretation in terms of self-serving biases, Skowronski, Betz, Thompson, and Shannon (1991) found an enhanced recall for pleasant, as opposed to unpleasant, events recorded in a diary related to the self, except for

events whose occurrence was considered to be highly typical or highly atypical. In other words, an example of Taylor's (1991) minimization stage comes from studies on self and memory. These include the works on self-enhancement showing individuals' tendency to recall fewer negative behaviors central to their self-conception (i.e., behaviors related to untrustworthiness and unkindness) when self-referenced compared to positive ones (i.e., behavioral instances of trustworthiness and kindness) and to those referred to another person in order to protect themselves against threats to the self (Alicke & Sedikides, 2009; Sedikides & Green, 2000, 2004; Sedikides & Gregg, 2008).

Affective Accounts. An example of the compatibility of the cognitive and affective accounts of the negativity effect comes from Peeters and Czapiński's (1990) behavioral-adaptive theory. The negativity effect would be part of a positive-negative asymmetry, and it would be functional to avoiding aversive stimuli and to limit the risks of an approach tendency toward stimuli that could be detrimental if generalized to all, and not only positive, stimuli. The behavioral-adaptive theory can explain the same asymmetries in the morality versus competence domains accounted for by cognitive theories, such as the schematic model and the cue-diagnosticity account. Peeters and colleagues (e.g., Peeters, 1989) found a negativity effect only when an "other-profitable" evaluative dimension was involved. This dimension encompasses attributes whose perceived adaptive relevance relates to *others* as opposed to the *self* (self-profitable dimension). The negativity effects have mostly been found when using paradigms eliciting judgments on either the likeability or the morality of a person (e.g., Kanouse & Hanson, 1972; Peeters & Czapiński, 1990), which are related to the approachability of a person for others. Where judgments of a person's competence, thus related to self-profitability, are concerned, no negativity effects are found (Peeters & Czapiński, 1990).

Aside from the "affective negativity effect" along the approach (like)–avoidance (dislike) continuum of behaviors described by Peeters and Czapiński (1990), there are other affective processes, such as some social perceiver's reactions to violations of morality, that could also contribute to explaining the negativity effect. Trafimow and colleagues (2005) found a causal relationship between negative affect associated with violations of morality and trait attribution. For example, in their Study 3, negative affect (induced, for example, by showing two video clips from the movie *Full Metal Jacket*) led participants to require fewer instances of violations of imperfect duties (e.g., unfriendly behaviors) to disconfirm a positive expectation (e.g., that the target is a friendly person) compared to when negative affect was not induced. The authors also showed that the effect of negative affect on attribution was produced by a range of emotions, including disgust, sadness, and fear. Thus, the authors proposed that affect could largely, although not solely, explain the weight people assign to immoral behaviors (Trafimow et al., 2005).

Finally, the social perceiver's reactions to threatening stimuli could account for the enhanced weight to the negative information on a target person's morality. Studies on social hypothesis testing and impression formation have highlighted the role of perceived threat in the negativity effect in the morality domain. People

tend to look for information that can falsify the presence of morality-related traits (e.g., honest), but not sociability-related (e.g., friendly) and competence-related (e.g., intelligent) traits, in a target person (Brambilla et al., 2011). The authors' socio-functionalist interpretation in terms of a protecting information-search strategy to avoid the risks of potentially harmful, immoral behaviors was corroborated by a subsequent study. In that study, Brambilla and co-workers (2012) found that the relationship between morality-based descriptions of a fictitious ethnic group (the Ortandesi) and group global impressions (measured in terms of their feelings of affection, hostility, hatred, and suspicion) was mediated by perceived threat (measured by the following items: "the Ortandesi pose a threat to Italian citizens," "the Ortandesi pose a threat to Italian values and beliefs," "the Ortandesi are dangerous for the stability of Italian economic system," "the Ortandesi threaten the Italian culture"; Brambilla et al., 2012, Study 3).

THE INTERPLAY BETWEEN COGNITIVE AND MOTIVATIONAL/AFFECTIVE FACTORS

As previously discussed, the motivational and affective explanations are not in contradiction with the cognitive accounts of the negativity effect. In contrast, there is an interplay between cognitive and motivational/affective factors underlying the negativity effect in the morality domain.

We have previously reviewed studies on self-enhancement and memory distortion, whereby individuals' memory for self-referenced behavioral instances that would threaten core aspects of their self, such as immoral behaviors (e.g., related to untrustworthiness), are recalled worse than positive exemplifications of those traits, as well as compared to memory of other people's central negative behaviors (Alicke & Sedikides, 2009; Sedikides & Green, 2000, 2004; Sedikides & Gregg, 2008). These studies indicate how motivations to protect the self can bias cognitive processes, such as memory. Additional evidence of this motivational-cognitive interplay comes from the studies on autobiographical memories by Walker and colleagues (1997) showing the enhanced emotional intensity associated with pleasant compared to unpleasant events over time (the fading affect bias), and by Skowronski and colleagues (1991) indicating enhanced recall for pleasant, as opposed to unpleasant, events related to the self as a function of their typicality in relation to the target.

These studies are not the only ones to provide evidence for the compatibility of the analysis of cognitive processes, such as memory, with a socio-motivational framework to increase our understanding of the negativity effect. For instance, the role of an actor's motives has been the focus of the multiple inference model proposed by Reeder and colleagues (Reeder, 2009; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002; Reeder, Vonk, Ronk, Ham, & Lawrence, 2004). This model further specifies the schematic model, which was based on how the social perceivers use implicit assumptions about trait-behavior relations to adjust their initial dispositional inferences (Reeder, 1993; Reeder et al., 1982). For example, an immoral behavior, such as stealing money from a charity (Reeder & Spores, 1983), would

lead to a behavior-correspondent trait inference of immorality regardless of situational demands, such as having been requested to steal the money. In contrast, making a donation would lead to a weaker behavior-correspondent trait inference of morality when the situation facilitated the behavior, such as when someone requested to make the donation (Reeder & Spores, 1983). The multiple inference model explains the negativity effect in attributions of morality by considering the consistency between the perceived actor's motives and the behavior-correspondent trait (Reeder, 2009). For example, if the observer infers that the actor is donating money to ingratiate the person who requested the donation, then this motive would be perceived as inconsistent with the presence of high levels of a morality-related trait, and it would lead to a weaker behavior-correspondent attribution of morality. Vice versa, the ingratiation motive would be perceived as consistent with the actor's low morality in the case of stealing money, and thus the behavior-correspondent inference of immorality would be made (Reeder, 2009).

Another example of the motivational/affective and cognitive interplay in producing the negativity effect comes from research showing that people's attention to negative morality information about a person is so high that they are not motivated to pay attention to additional information they receive about that person's competence (De Bruin & Van Lange, 2000, Study 2).

Recent work by Rusconi and colleagues (2017) has shown a reversal of the negativity effect in the morality domain when perceivers are asked to judge the likelihood or frequency of trait-inconsistent behaviors for moderate levels of both traits (e.g., "sincere") and behaviors (e.g., "covering for somebody"). The observed tendency to question a person's morality by assuming that she/he would more likely behave inconsistently than an immoral person was interpreted by the authors with reference to a socio-functionalist perspective, whereby perceivers are motivated to avoid the risk of omitting the detection of potential sources of threat (e.g., immoral behaviors).

Further, emotional responses are linked to and might even encompass cognitive variables. For example, frequency might underlie the affective reactions to immoral behaviors. The emotions manipulated by Trafimow and colleagues (2005) in their study on the causal role of affect in the negativity effect in the morality domain differ in terms of how often they occur in daily life. For instance, a study by Myrtek (2004) found that disgust was reported less frequently than sadness and anxiety/fear. In addition, emotions are not necessarily at odds with cognition, given the tight link between the two (Frijda, Manstead, & Bem, 2000) and the cognitive components (e.g., appraisals) involved in emotions (e.g., Cacioppo & Gardner, 1999; Trafimow et al., 2005).

THE BOUNDARY CONDITIONS OF THE NEGATIVITY EFFECT

In this section, we will review the main moderators of the negativity effect that emerged from the research conducted in the last four decades, as the literature has not yet provided a systematization. As previously discussed, a key and long-known moderator of the negativity effect in evaluative processes is the self-profitable

(competence-related) versus other-profitable (morality-related) dimension (e.g., Peeters & Czapiński, 1990; Reeder et al., 1982; Skowronski & Carlston, 1987, 1989). That is why several studies have focused their analysis of the negativity effect on the morality domain (e.g., Skowronski, 2002; Trafimow et al., 2005). However, research in the last few decades has pointed out some additional boundary conditions to the negativity effect. The relevance of examining boundary conditions lies in the opportunity they give scholars to cast light on the limits of existing theories, thus fostering their refinement or the development of new theories to enhance our understanding of the mechanisms underlying this phenomenon.

CONSISTENCY OF THE BEHAVIOR SET

According to the cue-diagnosticity account, diagnosticity has a role only when there are alternative categories to choose from. If the social perceiver examines a single behavior or a homogeneous set of behaviors (e.g., a series of failures) that points to the same trait category (e.g., incompetence), then other mechanisms should kick in. More specifically, Skowronski (2002; for a similar account, see Lupfer, Weeks, & Dupuis, 2000) proposed that when a single trait category is implied, judgments should be guided by a heuristic (a mental shortcut): Representativeness, that is, a judgment based on how similar a behavior is to a standard (e.g., Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974). A similar strategy has been described in the hypothesis-testing literature: Positive testing is an information-search strategy that focuses on what is more likely under a working hypothesis, such as "incompetence," than an alternative hypothesis, such as "competence" (e.g., Cherubini, Rusconi, Russo, Di Bari, & Sacchi, 2010; Klayman, 1995; Klayman & Ha, 1987; McKenzie, 2004). Positive testing and representativeness are akin to the strategy described by Hogarth and Einhorn (1992) for evaluation tasks, whereby people encode evidence with reference to their current hypothesis (Hogarth & Einhorn, 1992, p. 9).

TYPE OF MEASUREMENT OF TRAIT-BEHAVIOR RELATIONS

Trait-behavior relations represent a relatively neglected area in attribution research (Reeder, 1993; Reeder et al., 1982). This relative neglect has generated inconsistencies in the way traits, assumptions about traits and behaviors, and the negativity effect are measured in the judgment, impression formation, and attribution literatures. A consideration of the different types of measurement of trait-behavior relations can foster a reconciliation of some of the inconsistencies in the literature that we review here.

Reeder and colleagues (1982) analyzed the factors influencing the implicit assumptions about trait-behavior relations described in the schematic model by Reeder and Brewer (1979). More specifically, they analyzed three determinants of the attribution schemes: The social perceiver's considerations about central tendency (i.e., the tendency to use trait adjectives that average the observed behaviors), the actor's competence, and the social desirability of the actor's behaviors.

To investigate these determinants and their interplay with trait content (morality, competence, and preference), they used three measures of trait-behavior relations based on Heider's (1958) analysis of action (Reeder et al., 1982). Heider's (1958) "naïve analysis of action" focused on the personal and environmental forces underlying an actor's action. The personal force is determined by a power factor (often related to competence: *can*) and a motivational factor (*trying*). In line with this view, Reeder and colleagues (1982) measured trait-behavior relations in terms of what individuals with a trait *can* do ("perceived potential variability"), operationalized by asking the observers questions such as: "To what extent do you think that a person who is very intelligent (unintelligent) can adequately portray a person who is very unintelligent (intelligent)?" (Reeder et al., 1982, p. 361). In addition, they tested trait-behavior relations in terms of what the actors are perceived to *try* to do ("perceived intended variability"), operationalized as: "If a large reward were available for doing so, how likely is it that a person who is very honest (dishonest) would try to act very dishonest (honest)?" (Reeder et al., 1982, p. 361). Finally, Reeder and colleagues measured how *frequently* actors are perceived to emit behaviors ("perceived general variability"), operationalized as: "In general, how often does a very honest (dishonest) person act very dishonest (honest)?" (Reeder et al., 1982, p. 361). This latter measure has been the most influential in the literature, as subsequent works in this field have mostly used measures similar to the general variability one (Rusconi et al., 2017; Skowronski & Carlston, 1987, Experiment 1; Tausch, Kenworthy, & Hewstone, 2007, Study 3). Reeder and colleagues introduced the general variability measure alongside the intended variability measure to test trait-behavior relations along the morality dimension. The prediction was that both measures would elicit an asymmetry in line with Reeder and Brewer's (1979) model (Reeder et al., 1982, Table 1). Indeed, due to considerations of desirability, social perceivers would expect social actors to *try* to engage in moral behaviors rather than immoral ones. These intentions should then directly translate into behaviors. Thus, social perceivers would also assume that individuals would *more frequently* engage in moral, as opposed to immoral, behaviors (Reeder et al., 1982, p. 360). While the findings for the intended variability measure were in line with these predictions, the results for the general variability measure were not. In particular, in terms of behavior attempts (intended variability), target persons were thought to more likely try to behave in a socially desirable (e.g., honest, intelligent) manner than in a socially undesirable (e.g., dishonest, unintelligent) manner, more so for traits related to morality and competence compared to preference traits. However, in terms of general behavior frequency (general variability), the results were less clear-cut. In Experiment 1, target persons were deemed to more frequently engage in socially desirable (e.g., honest, intelligent) behaviors than in socially undesirable (e.g., dishonest, unintelligent) behaviors, as predicted. However, the pattern was similar for morality- and competence-related traits, contrary to the prediction of a greater effect for morality. In Experiment 2, there were no significant effects concerning the general variability measure.

Reeder and colleagues (1982) interpreted these unexpected and inconsistent findings yielded by the general variability measure in terms of one of the proposed determinants of trait-behavior relations, that is, the social perceiver's considerations of central tendency. Social perceivers would assume that it is very unlikely that a person with an extreme (either very moral or very immoral) trait would engage in behaviors at the opposite end of the trait dimension (either very immorally or very morally). This is because there is a wide discrepancy between the trait and the behavior in terms of extremity ("extreme distance"). The authors also proposed an alternative explanation in terms of a "reverse halo effect," whereby moral behaviors would be tainted and seen as immoral when they are emitted by immoral persons. These explanations could account for the floor effect found for morality-related traits at the extreme level of distance. However, they do not account for the inconsistencies between the results of Experiments 1 and 2, when there is a reduced discrepancy between a trait and a behavior in terms of level of extremity ("moderate distance"). In Experiment 1, an asymmetry toward socially desirable behaviors was found when considering people with moderate traits emitting extreme behaviors. In Experiment 2, no asymmetry was found when the relations between extreme traits and moderate behaviors were examined. In a similar way, central tendency cannot account for Rusconi and colleagues' (2017) findings. They used a measure similar to the general variability one, for example: "How often does an insincere (sincere) person tell the truth (omit some information)?" (Rusconi et al., 2017, p. 15). Their focus was on the observers' perceived trait-behavior relations when the level of distance between traits and behaviors is null, that is, when they are both moderate in terms of evaluative extremity. Across a series of studies using both abstract and concrete categories of behaviors and testing both Italian and American participants, they found a reversal of the negativity effect in morality along with a positivity effect in competence. In other words, participants judged it more likely that a person described as moral would more often behave in an immoral manner than the reverse (see Figure 2). Thus, Rusconi and colleagues' (2017) results indicate that social perceivers hold negative expectations about a moral person's behaviors. This finding dovetails with the results of the study by Meindl and colleagues (2016) who, moving from behavior assumptions to trait assumptions, have shown that these negative expectations characterize trait attributions as well (e.g., "Imagine you witness a stranger act immorally. How likely is it that they are an immoral *person*?" Meindl et al., 2016, p. 542). In particular, people are more ready to infer immorality from an immoral behavior than unsociability from a cold behavior ("the immoral assumption effect"; Meindl et al., 2016).

In sum, different measures of trait-behavior relations lead to different positive-negative asymmetries. The lack of a negativity effect when using the general variability measure for moderate behaviors (Reeder et al., 1982, Experiment 2) or its reversal when both traits and behaviors are moderate (Rusconi et al., 2017) also indicates the need to take into account another moderator of the negativity effect: The evaluative extremity of traits and behaviors.

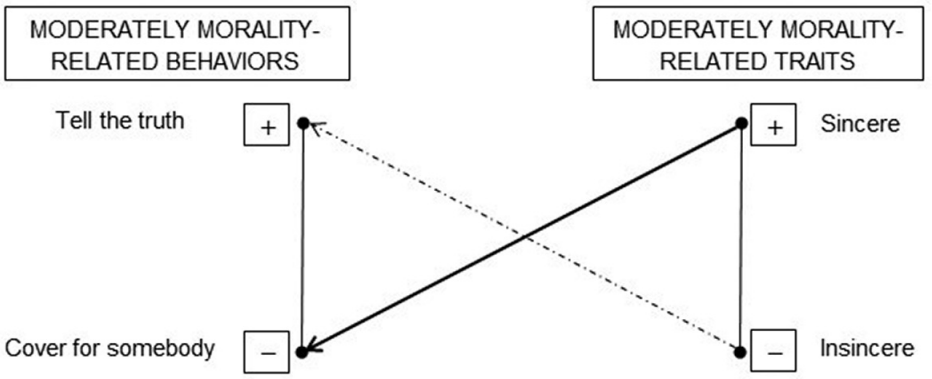

FIGURE 2. Representation of trait-behavior relations for moderate levels of positive and negative morality-related traits and behaviors according to Rusconi and colleagues (2017). + and - indicate traits/behaviors of positive and negative valence, respectively. The solid arrow indicates a strong trait-behavior relation, while the dashed arrow indicates a weak trait-behavior relation.

TRAIT EVALUATIVE EXTREMITY

Skowronski and Carlston's (1987, Experiment 1) study used a question phrasing that is similar to the one of the general variability measure ("Would an honest (dishonest) person ever search for the owner of a lost package?" Skowronski & Carlston, 1987, pp. 691-692; see Rusconi et al., 2017, S1 Text). It showed a positivity effect for the intelligent/stupid dimension and a negativity effect for the honest/dishonest dimension. These findings were interpreted in line with Reeder and Brewer's (1979) model. However, Skowronski and Carlston examined *moderate traits* (i.e., honest/dishonest and intelligent/stupid) in relation to different levels of extremity of the behaviors (i.e., extreme, moderate, and neutral). In contrast, Reeder and Brewer's model predicted the positivity and negativity effects mostly to account for relations between *extreme* traits and *extreme* behaviors, such as *very* dishonest people committing crimes (see also Rusconi et al., 2017, for a discussion).

In a manner similar to the one used in the Skowronski and Carlston (1987, Experiment 1) study, Tausch and colleagues (2007) used *moderate traits* in their extension of Rothbart and Park's work on the (dis-)confirmability of traits. In their Study 3, they used a measure of the likelihood of trait-inconsistent behaviors that was similar to Skowronski and Carlston's (1987, Experiment 1) dependent variable and Reeder and colleagues' (1982) measure of general variability ("how likely it is that someone who possesses the given trait shows trait-inconsistent behaviors [*diagnosticity*]?" Tausch et al., 2007, p. 550; see Rusconi et al., 2017, S1 Text). They tested this measure with moderate warmth- and competence-related traits (e.g., "trustworthy," "tolerant," "intelligent," "incompetent"). Their results showed a positivity effect in the competence domain and a negativity effect in the warmth

domain. These findings are in keeping with Skowronski and Carlston's (1987, Experiment 1), and they were interpreted as consistent with Reeder and Brewer's (1979) model, which, however, as explained above, predicted positivity and negativity effects mostly for *extreme* levels of both traits and behaviors.

In other words, both Skowronski and Carlston (1987, Experiment 1) and Tausch and colleagues (2007, Study 3) found positivity and negativity effects along the competence and morality dimensions, respectively, despite the fact that they used *moderate* traits. This could suggest that the level of extremity of traits is not crucial to produce positivity and negativity effects. However, there is not much systematic research on how *trait extremity* (e.g., "honest/dishonest" vs. "*very* honest/*very* dishonest") affects trait relations to behaviors at different levels of extremity (for an exception, see the study by Reeder et al., 1982, previously described). In particular, whether the use of extreme traits would elicit positivity and negativity effects of the same magnitude as those found with moderate traits or stronger effects is an open question.

BEHAVIOR EVALUATIVE EXTREMITY

As explained above, Skowronski and Carlston (1987) took into account *behavior extremity*. They found that both morality and competence judgments were characterized by an extremity bias, whereby impressions were more extreme than expected if the behaviors they were based on were averaged. In other words, extreme behaviors were weighed more than moderate behaviors (Skowronski & Carlston, 1987, Experiment 2). Their data also showed that the more extreme a behavior is, the more diagnostic it is (Skowronski & Carlston, 1987, Figures 3–4). In a subsequent study, Skowronski and Carlston found that trait inferences from extremely immoral behaviors and extremely intelligent behaviors were more resistant to change following contradictory evidence than trait inferences from behaviors of opposite valence (Skowronski & Carlston, 1992; for a similar finding concerning impressions of morality, see Reeder & Coovert, 1986; see also Mende-Siedlecki et al., 2013). As the authors noted, the finding about the "intelligent/stupid" trait category (but not the one for the "honest/dishonest" dimension) is at odds with Rothbart and Park's (1986) results on the confirmation of positive traits, such as "honest" and "intelligent," that are more difficult to acquire and require more instances to be confirmed than their disconfirmation does (Rothbart & Park, 1986, Table 9). They are also inconsistent with Tausch and colleagues' (2007, Study 2) results of a lack of significant difference between the number of instances required to disconfirm positive versus negative competence-related traits. Skowronski and Carlston argued that an explanation of this inconsistency might lie in the different methodologies used. Rothbart and Park (as well as Tausch et al.) analyzed the (dis)confirmability of *traits* without presenting any behaviors, thus their participants might have thought of the typical, *moderate* intelligent behaviors encountered in their everyday life. In contrast, Skowronski and Carlston presented *extreme* behaviors to their participants (Skowronski & Carlston, 1992, pp. 448-449). Thus, Skowronski and Carlston argued that the level of extremity

of the behaviors on which the trait inferences are based moderates the positivity effect in the competence domain, which should be stronger when social perceivers encounter extreme behaviors.

This account runs counter to Wojciszke, Brycz, and Borkenau's (1993) predictions of a stronger positivity effect for moderate behaviors in the competence domain and a stronger negativity effect for extreme behaviors in the morality domain. Their predictions rely on an approach-avoidance explanation, whereby negative, especially extremely negative, behaviors should elicit avoidance and a negativity effect because of their harmful consequences, whereas behaviors that are evaluatively moderate should induce approach and a positivity effect because they are considered less risky and their consequences more reversible. The results of their study were consistent with their hypotheses. They found that trait inferences and global evaluations of targets were affected by a stronger positivity effect in the competence domain and a lack of the negativity effect in the morality domain for moderate levels of traits and behaviors. These results echo those by Czapiński (1986) on stimulus traits' intensity, whereby the strongest negativity effect (the greatest diagnosticity) was found for the moderately intense features used to discriminate among target peers (see Peeters & Czapiński, 1990, Figure 2.2).

Thus, although evidence interpreted within the cue-diagnosticity framework (Skowronski & Carlston, 1992) and results interpreted within a motivational-based account (Wojciszke et al., 1993) agree on the presence of a negativity effect in the morality domain when the available behaviors are evaluatively extreme, they disagree on the level of behavior extremity that elicits the positivity effect in the competence domain. According to Skowronski and Carlston (1992), the positivity effect is stronger for extreme behaviors, whereas for Wojciszke and colleagues (1993), it is more evident when the available information is evaluatively moderate.

In keeping with the lack of negativity effect for moderately moral traits and behaviors found by Wojciszke and colleagues (1993), as previously described, Rusconi and colleagues (2017) found that participants assumed that a moderately moral target person was more likely to engage in moderately trait-inconsistent behaviors than a moderately immoral target person was. Future research should investigate trait-behavior relations at moderate levels of both traits and behaviors (null extremity distance) related to morality by using the intended variability measure, which was specifically introduced to test the morality dimension (Reeder et al., 1982).

LEVEL OF CATEGORIZATION: THE MODERATION OF PERCEIVED TARGET ENTITATIVITY

The schematic model by Reeder and Brewer (1979) argues that the social perceiver interprets an individual's behaviors as an organized pattern rather than a random selection from all the possible actions. In this sense, the schematic model echoes Asch's (1946) analysis of the configural model of impression formation, whereby the different traits that characterize a person are perceived as interrelated and they thus contribute to a unitary impression of a person (Asch, 1946). It follows that the

schematic model predicts a negativity effect in the morality domain for judgments involving a "unit formation," that is, for target individuals or groups made up of individuals related to one another, but not for aggregates of individuals unrelated to each other (Coovert & Reeder, 1990; Reeder & Coovert, 1986). Coovert and Reeder (1990, Experiment 1) confirmed this prediction. They found a negativity effect in morality impressions when the target was a person (described by two different, either moral or immoral, behaviors) or a pair of two friends (each described by one behavior), but not when the target was a pair of unrelated individuals (each described again by one behavior). Skowronski (2002) also examined the negativity and positivity effects at the individual versus group level. In a series of experiments, he presented participants with sets of consistent/inconsistent behaviors related to honesty/dishonesty and intelligence/unintelligence performed by either individuals or social club members or family members. Participants were then asked to make utility, risk/reward, consistency, evaluative, and trait judgments. He found that both the negativity and positivity effects were reduced in judgments of groups (i.e., families and social clubs). This finding is expected from a diagnosticity perspective given the greater variability (and amount) of behaviors that we could expect from a member of a group as opposed to an individual. For example, a member of an honest group should be perceived as more likely to act dishonestly than an honest individual should (Skowronski, 2002, p. 139). Although families fit with Coovert and Reeder's definition of "unit formation" because they represent a "meaningful group" rather than a mere "aggregate" (Coovert & Reeder, 1990), and thus impressions of families should be affected by the negativity bias, it should be noted that Coovert and Reeder (1990) also found that the negativity effect was strongest when the target was a person rather than a meaningful group.

In line with these results, the study by Welbourne (1999) showed greater positivity and negativity effects for individuals than groups in an impression-formation task. More specifically, individual targets were rated lower on kindness than group targets when participants formed impressions based on inconsistent (a mix of kind, unkind, and irrelevant) behaviors. Vice versa, intelligence ratings were more positive for individuals than for groups. Participants were also asked to rate the expected and actual consistency in behaviors of individual versus group targets. This measure was used to capture the perceived entitativity of individuals and groups. The results indicated that participants were more influenced by the behavior inconsistency of individuals rather than groups, thus suggesting that they expected a greater behavioral unity from individuals than groups (Welbourne, 1999, Study 1). However, the greater entitativity of individual versus groups, operationalized as behavioral consistency, did not account for the differences in morality and competence ratings between individuals and groups when experimentally manipulated (Welbourne, 1999, Study 2). In a third study, Welbourne (1999) operationalized entitativity in a different way, as a unity of target's intentions and goals. This study revealed a greater negativity effect in the morality domain for (either individual or group) targets high in entitativity as opposed to low-entitativity targets (in the competence domain, there was only a trend toward

positivity). According to the author, these findings show that only when the social perceivers expect that the target's behaviors are underlain by unified intentions and goals (thus indicating high entitativity) will they try and resolve any perceived behavioral inconsistency of the target by applying the attribution schemas based on diagnosticity. These schemas yield the negativity and positivity effects in the morality and competence domains, respectively.

According to Skowronski (2002), the reduced negativity effect for groups, as opposed to individuals, is not due to the differential processing of behavioral inconsistency for high- versus low-entitativity targets. Evidence against the role of behavioral consistency in driving the positivity and negativity effects came from his Experiment 2. Participants were asked to make a series of judgments about a target (an individual, a social club, or a family member, depending on the experimental condition) based on a set of either trait-consistent or trait-inconsistent behaviors. One of the requested judgments was about the consistency of the target's behaviors in a set with one another. The finding that individuals (as opposed to social club and family members) were not always perceived as the more behaviorally inconsistent targets in the case of trait-inconsistent behavior sets is in contrast with Welbourne's (1999) predictions of a greater effect of individuals' behavioral inconsistency as opposed to groups' behavioral inconsistency. In a similar way, the perceived consistency of individuals' behaviors was not the greatest in the case of trait-consistent behavior sets, contrary to what we should expect based on Welbourne (1999). In addition, based on Welbourne (1999), the perceived inconsistency judgments should mediate trait judgments, in particular the reduced positivity and negativity effects for judgments about groups, as opposed to individuals. However, the mediational analysis did not provide support for the mediation role of consistency judgments (Skowronski, 2002, Experiment 2). Instead, data supported the predictions of the cue-diagnosticity account based on the differential diagnosticity of positive and negative behaviors.

Altogether, these studies consistently point to the moderating role of the target's level of categorization, and in particular the perceived unity underlying the target's actions (target entitativity), on the negativity and positivity effects. Cognitive theories based on the informativeness of behaviors, such as the schematic model and the cue-diagnosticity account, can account for the differences due to target entitativity as shown by the above reviewed studies by Coovert and Reeder (1990) and Skowronski (2002, Experiment 2), respectively. In contrast, alternative models cannot fully account for the moderation of perceived target entitativity. The predictions based on the processing of inconsistency in behaviors (Welbourne, 1999) and motivation- and affect-based theories (Ito, Larsen, et al., 1998; Peeters & Czapiński, 1990) did not receive empirical support (Skowronski, 2002, Experiment 2). Finally, the novelty approach (Fiske, 1980), the correspondent inference theory (Jones & Davis, 1965; Jones & McGillis, 1976), expectancy-contrast theories (Helson, 1947, 1948; Sherif & Sherif, 1967), and the range theories (Birnbaum, 1972; Wyer, 1974) cannot account for target entitativity because they focus on the properties of stimuli, such as their novelty (Fiske, 1980), ambiguity (Birnbaum, 1972;

Wyer, 1974), and relationship with the social perceiver's anchors (Helson, 1947, 1948; Sherif & Sherif, 1967) and constructs (Jones & Davis, 1965; Jones & McGillis, 1976), rather than the organization underlying the target's actions.

BELIEFS IN PERSONALITY STABILITY: ENTITY VERSUS INCREMENTAL THEORISTS

An individual difference in social perception and interaction is the belief that people have in lay theories, such as those about the "static," as opposed to "dynamic," view of human beings and the world (Levy, Plaks, Hong, Chiu, & Dweck, 2001). For example, people can believe that intelligence- and morality-related attributes are fixed ("entity theory") or malleable ("incremental theory"; Dweck, Chiu, Hong, & Inquiry, 1995). Skowronski's (2002) Experiment 2 tested people's individual difference in beliefs about personality stability. The motivation- and affect-based theories predict that social perceivers who believe in personality as fixed (entity theorists) and social perceivers who believe in personality as malleable (incremental theorists) are differently influenced by stimuli, and thus the trait ratings based on consistent (a homogeneous set of honest behaviors) or inconsistent (a set of mostly honest behaviors and one dishonest behavior) behaviors should also differ across the two types of theorists. In contrast, accounts based on behavior informativeness, in particular the cue-diagnosticity account, predict that entity theorists should discount the informativeness of a single, inconsistent behavior within a set of otherwise consistent behaviors. Incremental theorists should instead be more influenced by a single, inconsistent behavior. In addition, according to the cue-diagnosticity account, no differences between entity and incremental theorists should be observed when the set of behaviors the judgments are based on is consistent. Indeed, as previously discussed, heuristic judgments rather than diagnosticity considerations would be used when the examined behaviors imply a single category. The results showed that entity theorists were not influenced in their honesty and intelligence trait ratings by a single inconsistent behavior within a set of behaviors that consistently implied the same single trait (i.e., either honest, dishonest, intelligent, or unintelligent) to the same extent as incremental theorists were. In other words, entity theorists were more influenced in their trait judgments by the valence of the majority of the behaviors in a set than were incremental theorists. These results are in line with the accounts relying on the informativeness of behaviors, such as the schematic model and the cue-diagnosticity account. However, they cannot be fully accounted for by motivation- and affect-based theories because they predict different patterns of responses in entity versus incremental theorists. In contrast, the pattern of results showed only a different polarization in entity versus incremental theorists' trait ratings (see Skowronski, 2002, Figure 3).

No difference between entity theorists and incremental theorists was found when the trait judgments were based on a consistent set of behaviors. This lack of difference runs counter to the difference predicted by the motivation- and affect-based theories. However, it is in line with the prediction of the cue-diagnosticity

account, whereby social perceivers should rely on heuristic judgments, such as representativeness, rather than on diagnosticity considerations when only one trait category is implied by a set of behaviors.

SELF AND THE CLOSENESS OF THE RELATIONSHIP WITH THE TARGET

As previously reported, self-enhancement can moderate the negativity effect because individuals tend to exhibit worse recall of self-referenced negative behaviors central to the self, such as those pointing to their untrustworthiness (Alicke & Sedikides, 2009; Sedikides & Green, 2000, 2004; Sedikides & Gregg, 2008). In addition, Skowronski and Carlston (1987, Experiment 2) found that negative behaviors used in an impression-formation task about 25 different, unknown target persons were recalled better than positive behaviors (with the exception of moderately negative morality-related behaviors). This recall facilitation did not emerge in the diary study by Skowronski and colleagues (1991) described in a previous section, in which the actors of the behaviors were people close to the participants (e.g., their close friends). Skowronski and colleagues suggested that a close relationship between the observer and the actor might enhance the recall of the actor's positive behaviors. This suggests that the closeness of the observer-actor relationship can moderate the negativity effect in memory recall.

CULTURE AND HISTORICAL PERIOD

Unlike previous models' assumption that social perceivers hold moderately positive expectations about other people's behaviors (Fiske, 1980; Helson, 1947, 1948; Jones & Davis, 1965; Sherif & Sherif, 1967), the recent studies by Meindl and colleagues (2016) and Rusconi and colleagues (2017) point to a "cynical" view of social perceivers who assume that even moderately moral people could engage in moderately immoral behaviors (e.g., Rusconi et al., 2017). A possible interpretation of this discrepancy is that the perceivers' assumptions about other people's morality vary across culture and time. A systematic analysis of the influence of cultural factors is needed as research in this area has been conducted in Western countries (e.g., Reeder et al., 1982; Rusconi et al., 2017; Skowronski & Carlston, 1987, 1992; Tausch et al., 2007). Rusconi and colleagues (2017) found the same pattern of results, although with a different magnitude, in an Italian and an American sample (Rusconi et al., 2017, Study 4). It is thus possible that more pronounced differences can be found when comparing Western as opposed to non-Western participants, given the different consideration of situational information (e.g., Morris & Peng, 1994) as well as conceptualizations of morality (e.g., Shweder, Mahapatra, & Miller, 1987; Wojciszke, Bazinska, & Jaworski, 1998, footnote 2) across societies.

Time is another variable that could explain the differences in perceivers' assumptions about other people's morality. In the United States, the public's trust and confidence in politicians and American people has declined from the early 1970s

to 2018, but with fluctuations (Jones & Saad, 2018). However, the public's trust in professions among British adults aged 15 and over has been relatively stable between 1983 and 2017 (Ipsos MORI, 2017). In addition, it should be noted that the results of the studies conducted in this field with respect to the presence of a negativity effect in the morality domain have been relatively consistent between the 1970s and the 2010s (Reeder et al., 1982; Reeder & Spores, 1983; Skowronski & Carlston, 1987; Tausch et al., 2007). Thus, more empirical evidence to support the hypothesis of variations in culture and time at the basis of the shift in perceivers' expectations is needed.

The theoretical implication highlighted by this review of the main moderators of the negativity effect is that the properties of traits, correspondent and non-correspondent behaviors, and their relations play an important role in determining the negativity effect. However, the variables related to how the observers perceive the actors in terms of their relational closeness as well as their goals and intentions also play a role in determining the enhanced weight for negative events.

THEORETICAL AND EMPIRICAL INCONSISTENCIES: A TRAJECTORY FOR THE FIELD

As emerged from the previous analysis of moderators of the negativity effect, there are some theoretical and empirical inconsistencies in this research field. In this section, we will address them and outline the trajectory for this research area.

FREQUENCY AND DIAGNOSTICITY

There has been confusion between the constructs of diagnosticity and frequency in the literature. The term *diagnosticity* has been used with reference to different constructs by different authors (Mende-Siedlecki et al., 2013; Sanbonmatsu et al., 2015; Skowronski, 2002; Skowronski & Carlston, 1987, 1992). As argued by Rozin and Royzman (2001), diagnosticity and frequency correlate with one another, but they can be distinguished. For example, the difference between these two constructs has been empirically pointed out in the social hypothesis-testing literature. Rusconi, Sacchi, Toscano, and Cherubini (2012) asked participants to select some questions from a pre-set list to investigate the presence of a series of personality traits in an anonymous target person. They were also asked whether they expected to receive an answer confirming the tested trait or a disconfirming answer as well as the probability of receiving it. Participants exhibited a differential pattern of confirming expectations as a function of answer frequency (defined as likelihood to occur) and answer diagnosticity (defined as informativeness from a Bayesian perspective). The confirming expectations were higher when the confirming answers were moderately diagnostic *and* frequent compared to when there was a diagnosticity/frequency trade-off (i.e., confirming answers were either highly diagnostic but rare or vice versa). These results show that perceivers are sensitive to the congruency or incongruency of cue diagnosticity and cue frequency.

Skowronski and Carlston (1987) drew the concept of diagnosticity in impression formation from the categorization literature. The diagnosticity of a cue depends on its ability to predict a trait category; thus, even a frequent cue can be diagnostic if it leads to an accurate trait categorization. However, in the Mende-Siedlecki and colleagues' (2013) study, diagnosticity has been used with reference to a different conceptualization of informativeness. According to the authors, cue diagnosticity is an emergent property of cue frequency, and thus a rare cue is more informative than a common cue. Rather than diagnosticity as defined in the category learning theory, the latter conceptualization refers more squarely to the constructs of "surprisal" (Tribus, 1961), "self-information" in information theory (Rusconi, Crippa, Russo, & Cherubini, 2012), and the "rarity assumption" in the Bayesian reasoning literature (Anderson, 1990; McKenzie, 2006; McKenzie & Chase, 2012; McKenzie & Mikkelsen, 2000, 2007; Oaksford & Chater, 1994; Rusconi & McKenzie, 2013), whereby "answers, or test outcomes, are diagnostic to the extent that they are rare or surprising" (McKenzie, 2006, p. 580). Future work should differentiate the terminology used to define the constructs of diagnosticity as emerged from the categorization literature and diagnosticity as defined in the information-theory field.

In addition, a gap in recent research on behavior frequency (Mende-Siedlecki et al., 2013; Sanbonmatsu et al., 2015) is that it does not provide a direct comparison of the rarity account against any other explanatory mechanisms. Thus, for example, there is a lack of systematic investigations on the relation between categorization diagnosticity and information-theory diagnosticity that pits them against one another and clarifies the way they relate to cue frequency in light of people's sensitivity to and implicit assumptions about the rarity of events (McKenzie & Chase, 2012; Oaksford & Chater, 1994).

In their review of the negativity effect in impression formation, Rozin and Royzman (2001, p. 308) pointed out that the frequency and diagnosticity accounts share the prediction of a positivity effect, whereby social perceivers would weigh a single, positive behavior (e.g., an exceptional performance on a test) more than some negative behaviors (e.g., some fails) when the trait to be inferred is positive and rare (e.g., high intelligence; see Skowronski & Carlston, 1992). They also argued that both the frequency and diagnosticity accounts lack explanatory power for the negativity effect in the morality domain. They claimed that extremely moral behaviors (e.g., saving lives) are as infrequent and diagnostic as extremely immoral acts (e.g., murders), despite the latter receiving enhanced weight (Rozin & Royzman, 2001, p. 310). Systematic, empirical investigations of these theorizations, for example, by orthogonally manipulating frequency and diagnosticity, should clarify their respective role in producing the negativity effects.

MEASURES OF TRAIT-BEHAVIOR RELATIONS AND LEVELS OF TRAIT/BEHAVIOR EVALUATIVE EXTREMITY

The inconsistencies of the results about the positivity and negativity effects at different levels of evaluative extremity could be resolved by taking into account the

interplay between the different measures of trait-behavior relations (potential, intended, and general variabilities) and the (moderate vs. extreme) levels of both traits and behaviors in the morality versus competence domains. The manipulation of the level of evaluative extremity of traits and behaviors would require a stricter definition of "moderate" and "extreme." A parameterization of the different levels of evaluative extremity based on the observers' self-reports and also their physiological reactions would reduce the statistical noise that the ambiguity of the labels such as "*very* honest" might have generated in previous studies in the literature. This could also provide a basis for more squarely testing the interplay between the cognitive and the motivational and affective variables underlying the negativity effect.

An additional measurement aspect that future research in the field could take into account is the trait scale used to assess people's attributions. Typically, the scales used in the literature to assess trait judgments employ a bipolar continuum encompassing both trait categories implicated by the behaviors on which the trait judgments are based (e.g., very dishonest to very honest). Although observers might tend to reinterpret a unipolar scale as a bipolar scale, there is evidence in the literature indicating that using unipolar scales can unveil effects that could cancel each other out when using bipolar scales (Gamblin, Banks, & Dean, 2019). As previously discussed, the cue-diagnosticity account predicts positivity and negativity effects only when more than a single trait category is implicated by the available evidence. If unipolar scales to judge the presence of a single trait category (e.g., absence of honesty to presence of honesty or absence of dishonesty to presence of dishonesty) were used, we should observe no, or, if anything, reduced positivity and negativity effects, which could provide evidence in favor of the informativeness-based models, such as the cue-diagnosticity account, while being problematic for affect-based accounts without any additional assumptions.[1]

THE ROLE OF AFFECT

In contrast with the results supporting the role of affective reactions in producing the negativity effect in the morality domain (Trafimow et al., 2005), Skowronski (2002) showed that evaluative (likability and goodness) judgments of targets did not mediate trait judgments, and thus affective reactions to targets were not the basis for trait ratings (Skowronski, 2002, Experiment 2). Also, there was no negativity effect in the evaluative judgments of targets based on honest versus dishonest behaviors when targets were described by a consistent set of behaviors. Instead, there was an equal polarization of the participants' evaluations of honesty and dishonesty. Further, both with inconsistent and consistent behavior sets, there was a mismatch between the negative intelligence ratings and non-negative evaluations of targets who either mostly or consistently behave in an unintelligent way. Similarly, evaluative reactions to behaviors did not suggest a role of affect in the

1. We thank John Skowronski for the suggestions reported in this subsection.

asymmetric influence of behaviors on judgments about morality and competence (Skowronski, 2002, Experiment 1).

A possible reason for these discrepant results on the role of affect lies in the way this construct was operationalized in the two studies. While Skowronski (2002) focused on evaluations of the goodness or likability of behaviors and targets, Trafimow and colleagues (2005) examined both valence (participants' positive/negative feelings) and specific emotions, such as disgust, sadness, and fear (Trafimow et al., 2005, Study 5). Future research should clarify these inconsistencies for example by orthogonally testing different operationalizations of affect (e.g., in terms of likability vs. emotions). In addition, future research could consider the differences between negativity effects that are affective (defense-based, e.g., fear that relates to the avoidance of negative stimuli) and those that are informational (orienting-based, to control the sources of negative stimuli; Peeters & Czapiński, 1990).

CONCLUSIONS

Although there is consensus around the notion that negative events are weighed more than positive ones in forming impressions of others and in trait attribution, there is no such agreement on the underlying mechanisms. Indeed, although the informativeness-based models such as the cue-diagnosticity account and the schematic model have received support and they are able to account for several results in the literature (e.g., Skowronski, 2002), there are still areas that deserve empirical investigation. An example is the relative contribution of cognitive and affective variables to the negativity effect in the morality domain (e.g., Skowronski, 2002; Trafimow et al., 2005). Our review shows that the cognitive and motivational- and affect-based accounts of the negativity effect are compatible and they can jointly contribute to a deeper understanding of this phenomenon (e.g., Alicke & Sedikides, 2009; De Bruin and Van Lange, 2000; Rusconi et al., 2017; Sedikides & Green, 2000, 2004; Sedikides & Gregg, 2008; Skowronski et al., 1991; Walker et al., 1997).

In addition, the boundary conditions to the negativity effect emerged throughout decades of research provide the basis for studies that could resolve some of the theoretical and empirical inconsistencies emerged in the literature. In addition to the role of trait content (morality vs. competence, e.g., Peeters & Czapiński, 1990; Skowronski & Carlston, 1987, 1989), the negativity effect is qualified by the consistency of the behaviors on which the judgment is based (Lupfer et al., 2000; Skowronski, 2002), the level of categorization and perceived target entitativity (Coovert & Reeder, 1990; Skowronski, 2002; Welbourne, 1999), the level of evaluative extremity of traits and behaviors (Rusconi et al., 2017; Wojciszke et al., 1993), beliefs in lay theories of personality (Skowronski, 2002), and self-enhancement and the relationship between the observer and the actor (e.g., Alicke & Sedikides, 2009; Sedikides & Green, 2000, 2004; Sedikides & Gregg, 2008; Skowronski et al., 1991; Skowronski & Carlston, 1987). Another moderator is the type of measurement

used to test trait-behavior relations (Reeder et al., 1982; Rusconi et al., 2017). The literature has focused on the perceived frequency of behaviors (Fiske, 1980; Mende-Siedlecki et al., 2013; Sanbonmatsu et al., 2015) and on associated measures such as the general variability one (Reeder et al., 1982; Rusconi et al., 2017; Skowronski & Carlston, 1987; Tausch et al., 2007). This leaves relatively untested the measures that should more clearly distinguish between judgments of morality (the intended variability measure) and competence (the potential variability measure; Reeder et al., 1982; Skowronski, 2002).

Reeder and Brewer's (1979) hierarchically restrictive schema applies mostly to extreme levels of *both* traits (e.g., "a ruthless con man"; Reeder & Brewer, 1979, p. 68) and behaviors (e.g., "embezzlement"; Reeder & Brewer, 1979, p. 68). Although *behavior extremity* as a moderator of trait attribution has been taken into account (Skowronski & Carlston, 1987; Wojciszke et al., 1993), there is scant, systematic research considering *trait extremity* in relation to *behavior extremity* (Reeder et al., 1982). However, a recent focus on the *moderate* levels of both traits and behaviors has revealed the social perceivers' cynical expectations about actors' frequency of emission of moral behaviors (e.g., Rusconi et al., 2017). These negative assumptions about other people's morality (see also Meindl, Johnson, & Graham, 2016) question the traditionally assumed moderately positive expectations of the social perceiver (Fiske, 1980; Helson, 1947, 1948; Jones & Davis, 1965; Sherif & Sherif, 1967) and deserve further investigation to unveil potential variations across cultures and historical periods.

Future research would also benefit from a parameterization of the levels of evaluative extremity of both traits and behaviors. More precise definitions of relevant constructs such as "affect" and "diagnosticity" could also help resolve some of the theoretical and empirical inconsistencies in the literature. An example, is the clarification of the relation between behavior frequency and diagnosticity and their respective roles in producing the negativity effect (Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1987).

Finally, the hierarchically restrictive schema has been the focus of several investigations, but the two other schemata proposed in Reeder and Brewer's (1979) schematic model still deserve a systematic empirical investigation, namely the "partially restrictive schema" (typical of traits such as "friendly"/"unfriendly," Figure 3) and the "fully restrictive schema" (applicable to traits such as "neat"/"sloppy," Figure 4).

As shown by our review, the motivational and affective bases of the negativity effect are intertwined with the cognitive variables and processes, thus an integrated approach is needed to foster our understanding of the negativity effect. Such an approach could clarify the role of negativity in moderating intergroup biases (Hewstone, Rubin, & Willis, 2002, pp. 586-587). It could also complement the contribution of studies in social perception and intergroup relations that rely on socio-functionalist accounts to explain the primary role of morality in impression formation and updating (e.g., Brambilla et al., 2013; van der Lee, Ellemers, Scheepers, & Rutjens, 2017).

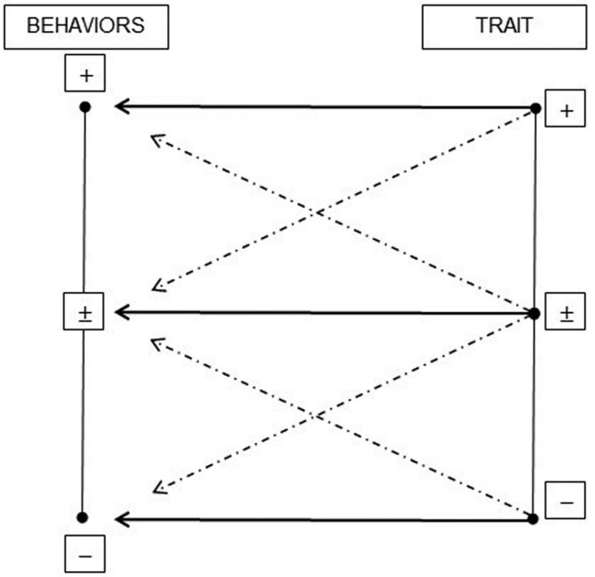

FIGURE 3. Representation of the trait-behavior relations according to the partially restrictive schema (modified from Reeder & Brewer, 1979, Figure 2). +, ± , and - indicate extremely positive, moderate, and extremely negative trait/behaviors, respectively. The solid arrow indicates a strong trait-behavior relation, while the dashed arrow indicates a weak trait-behavior relation. Without any specific information about the context this schema predicts moderate levels of behaviors that are not very informative about the implied traits. However, moderate behaviors can be informative about the underlying traits if there are situations that demand for extreme behaviors (Reeder & Brewer, 1979, Figure 3).

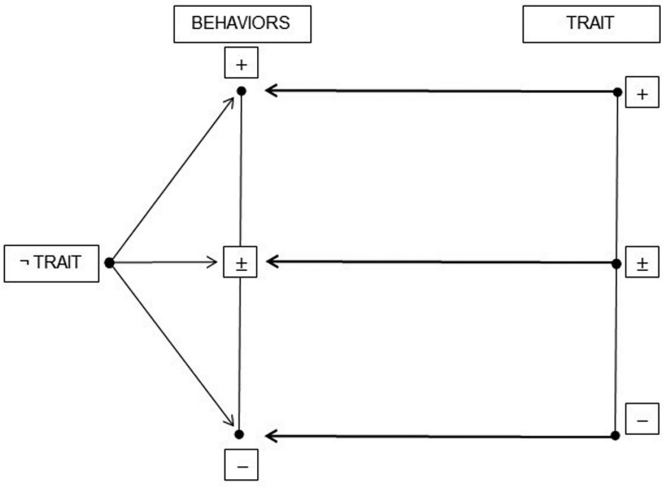

FIGURE 4. Representation of the trait-behavior relations according to the fully restrictive schema (modified from Reeder & Brewer, 1979, Figure 5). +, ± , and - indicate extremely positive, moderate, and extremely negative trait/behaviors, respectively. The thick solid arrows indicate the strong trait-behavior relation at all levels (extremely positive, neutral, and extremely negative) predicted by this schema, according to which people emit a restricted range of behaviors consistent with the trait level they possess. "¬ TRAIT" indicates the absence of a trait, which is implied whenever a person emits behaviors at all levels of the continuum in different circumstances.

REFERENCES

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400. https://doi.org/10.1037/h0022280

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*(3), 272–279. https://doi.org/10.1037/h0025907

Anderson, N. H., & Alexander, G. R. (1971). Choice test of the averaging hypothesis for information integration. *Cognitive Psychology*, *2*(3), 313–324. https://doi.org/10.1016/0010-0285(71)90017-X

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, *93*(1), 35–42. https://doi.org/10.1037/h0032589

Birnbaum, M. H. (1974). The nonadditivity of personality impressions. *Journal of Experimental Psychology*, *102*(3), 543–561. https://doi.org/10.1037/h0036014

Birnbaum, M. H., Parducci, A., & Gifford, R. K. (1971). Contextual effects in information integration. *Journal of Experimental Psychology*, *88*(2), 158–170. https://doi.org/10.1037/h0030880

Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology*, *78*(September 2017), 34–42. https://doi.org/10.1016/j.jesp.2018.04.011

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition*, *32*(4), 397–408. https://doi.org/10.1521/soco.2014.32.4.397

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology*, *49*(5), 811–821. https://doi.org/10.1016 /j.jesp.2013.04.005

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*(1), 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401–423. https://doi.org/10.1037//0033-2909.115.3.401

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual Review of Psychology*, *50*, 191–214.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*(1), 3–25.

Cherubini, P., Rusconi, P., Russo, S., Di Bari, S., & Sacchi, S. (2010). Preferences for different questions when testing hypotheses in an abstract task: Positivity does play a role, asymmetry does not. *Acta Psychologica*, *134*, 162–174. https://doi.org/10.1016/j.actpsy.2010.01.007

Coovert, M. D., & Reeder, G. D. (1990). Negativity effects in impression formation: The role of unit formation and schematic expectations. *Journal of Experimental Social Psychology*, *26*(1), 49–62. https://doi.org/10.1016/0022-1031(90)90061-P

Czapiński, J. (1986). Informativeness of evaluations in interpersonal communication: Effects of valence, extremity of evaluations, and ego-involvement of evaluator. *Polish Psychological Bulletin*, *17*(3–4), 155–164.

De Bruin, E. N. M., & Van Lange, P. A. M. (2000). What people look for in others: Influences of the perceiver and the perceived on information selection. *Personality and Social Psychology Bulletin*, *26*(2), 206–219. https://doi.org/10.1177/0146167200264007

Dweck, C. S., Chiu, C., Hong, Y., & Inquiry, S. P. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review 23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org /10.1037/0022-3514.38.6.889

Frijda, N. H., Manstead, A. S. R., & Bem, S. (2000). The influence of emotions on beliefs. In N. H. Frijda, A. S. R. Manstead, & S. Bem (Eds.), *Emotions and beliefs: How feelings influence thoughts* (pp. 1–9). New York: Cambridge University Press.

Gamblin, D. M., Banks, A. P., & Dean, P. J. A. (2019). Affective responses to coherence in high and low risk scenarios. *Cognition and Emotion*, *0*(0), 1–19. https://doi.org /10.1080/02699931.2019 .1640663

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org /10.1037/a0034726

Heider, F. (1958). The naive analysis of action. *Psychology of Interpersonal Relations*, 79–124. https://doi.org/10.1037/10628-004

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology*, *60*(1), 1–29. https://doi.org/10.2307/1417326

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review*, *55*(6), 297–313. https://doi.org/10.1037/h0056721

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*, 575–604.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010 -0285(92)90002-J

Ipsos MORI. (2017). Trust in professions: Long-term trends. https://ipsos.com/ipsos-mori/en-uk/trust-professions-long-term-trends

Ito, T. A., Cacioppo, J. T., & Lang, P. J. (1998). Eliciting affect using the International Affective Picture System: Trajectories through evaluative space. *Personality and Social Psychology Bulletin*, *24*(8), 855–879.

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*(4), 887–900. https://doi.org/10.1037/0022-3514.75.4.887

Jones, E. E. (1976). How do people perceive the causes of behavior? Experiments based on attribution theory offer some insights into how actors and observers differ in viewing the causal structure of their social world. *American Scientist*, *64*(3), 300–305.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 219–266). New York: Academic Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Jones, E. E., & McGillis, D. (1976). Correspondent inferences and the attribution cube: A comparative reappraisal. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New directions*

in attribution research (pp. 390–420). Hillsdale, NJ: Erlbaum.

Jones, J., & Saad, L. (2018). Gallup poll social series: Governance. https://news.gallup.com/file/poll/243428/181004trustpeopldpoliticians.pdf

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kanouse, D. E. (1984). Explaining negativity biases in evaluation and choice behavior: theory and research. *NA - Advances in Consumer Research*, *11*, 703–708.

Kanouse, D. E., & Hanson, L. R. J. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47–62). New York: General Learning Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. https://doi.org/10.1037/h0034225

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. Retrieved from file:%5C%5CC:%5CDocuments and Settings%5CMarkec%5CMyDocuments%5CFaks_Studije%5CElektroničkeKnjige%5CMarkovaZbirkaMudrolija.Data%5CPDF%5CKlaymanHa_HypothesesTesting-0027131136/KlaymanHa_HypothesesTesting.pdf

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart. *Personality and Social Psychology Bulletin*, *42*(9), 1272–1290. https://doi.org/10.1177/0146167216655984

Landy, J. F., Piazza, J., & Goodwin, G. P. (2018). Morality traits still dominate in forming impressions of others. *Proceedings of the National Academy of Sciences*, *115*(25), E5636. https://doi.org/10.1073/pnas.1807096115

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group Virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Levy, S. R., Plaks, J. E., Hong, Y. Y., Chiu, C. Y., & Dweck, C. S. (2001). Static versus dynamic theories and the perception of groups: Different routes to different destinations. *Personality and Social Psychology Review*, *5*(2), 156–168. https://doi.org/10.1207/S15327957PSPR0502_6

Lupfer, M. B., Weeks, M., & Dupuis, S. (2000). How pervasive is the negativity bias in judgments based on character appraisal? *Personality and Social Psychology Bulletin*, *26*(11), 1353–1366. https://doi.org/10.1177/0146167200263004

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200–219). Oxford, UK: Blackwell.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition*, *34*(3), 577–588. https://doi.org/10.3758/BF03193581

McKenzie, C. R. M., & Chase, V. M. (2012). *Why rare things are precious: How rarity benefits inference*. https://doi.org/10.1093/acprof

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review*, *7*(2), 360–366. https://doi.org/10.3758/BF03212994

McKenzie, C. R.M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004

Meindl, P., Johnson, K. M., & Graham, J. (2016). The immoral assumption effect: Moralization drives negative trait attributions. *Personality and Social Psychology Bulletin*, *42*(4), 540–553. https://doi.org/10.1177/0146167216636625

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Jour-*

nal of Neuroscience, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *67*(6), 949–971.

Myrtek, M. (2004). *Heart and emotion: Ambulatory monitoring studies in everyday life*. Göttingen, Germany: Hogrefe & Huber.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037//0033-295X.101.4.608

Peeters, G. (1989). Evaluative positive-negative assymetry in adjective-noun compounds. *Polish Psychological Bulletin*, *20*(4), 255–266.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, *1*(1), 33–60. https://doi.org/10.1080/14792779108401856

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391. https://doi.org/10.1037/0022-3514.61.3.380

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin*, *19*(5), 586–593. https://doi.org/10.1177/0146167293195010

Reeder, G. D. (1997). Dispositional inferences of ability: Content and process. *Journal of Experimental Social Psychology*, *33*(2), 171–189. https://doi.org/10.1006/jesp.1996.1316

Reeder, G. D. (2006). From trait-behavior relations to perceived motives: An evolving view of positivity and negativity effects in person perception. *Polish Psychological Bulletin*, *37*(4), 191–202.

Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry*, *20*(1), 1–18. https://doi.org/10.1080/10478400802615744

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

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

Reeder, G. D., Henderson, D. J., & Sullivan, J. J. (1982). From dispositions to behaviors: The flip side of attribution. *Journal of Research in Personality*, *16*(3), 355–375. https://doi.org/10.1016/0092-6566(82)90032-0

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology*, *83*(4), 789–803. https://doi.org/10.1037/0022-3514.83.4.789

Reeder, G. D., Pryor, J. B., & Wojciszke, B. (1992). Trait-behavior relations in social information processing. In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 37–57). London: Sage.

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*(4), 736–745. https://doi.org/10.1037/0022-3514.44.4.736

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology*, *86*(4), 530–544. https://doi.org/10.1037/0022-3514.86.4.530

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131–142. https://doi.org/10.1037/0022-3514.50.1.131

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957SPR0504_2

Rusconi, P., Crippa, F., Russo, S., & Cherubini, P. (2012). Moderators of the feature-positive effect in abstract hypothesis-evaluation tasks. *Canadian Journal of Experimental Psychology*, *66*(3), 181–192. https://doi.org/10.1037/a0028173

Rusconi, P., & McKenzie, C. R. M. (2013). Insensitivity and oversensitivity to an-

swer diagnosticity in hypothesis testing. *Quarterly Journal of Experimental Psychology*, *66*(12), 2443–2464. https://doi.org/10.1080/17470218.2013.793732

Rusconi, P., Sacchi, S., Capellini, R., Brambilla, M., & Cherubini, P. (2017). You are fair, but I expect you to also behave unfairly: Positive asymmetry in trait-behavior relations for moderate morality information. *PLoS ONE*, *12*(7), e0180686. https://doi.org/10.1371/journal.pone.0180686

Rusconi, P., Sacchi, S., Toscano, A., & Cherubini, P. (2012). Confirming expectations in asymmetric and symmetric social hypothesis testing. *Experimental Psychology*, *59*(5), 243–250. https://doi.org/10.1027/1618-3169/a000149

Sanbonmatsu, D. M., Mazur, D., Behrends, A. A., & Moore, S. M. (2015). The role of the baserate frequency of correspondent behavior and trait stereotypes in attribution: Building on Rothbart and Park (1986). *Social Cognition*, *33*(4), 255–283. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84941555578&partnerID=40&md5=05fe61308e558e317489f4727543b21a

Sedikides, C., & Green, J. D. (2000). On the self-protective nature of inconsistency-negativity management: Using the person memory paradigm to examine self-referent memory. *Journal of Personality and Social Psychology*, *79*(6), 906–922. https://doi.org/10.1037/0022-3514.79.6.906

Sedikides, C., & Green, J. D. (2004). What I don't recall can't hurt me: Information negativity versus information inconsistency as determinants of memorial self-defense. *Social Cognition*, *22*(1), 4–29. https://doi.org/10.1521/soco.22.1.4.30987

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science*, *3*(2), 102–116. https://doi.org/10.1111/j.1745-6916.2008.00068.x

Sherif, C. W., & Sherif, M. (1967). Attitudes as the individual's own categories: The social judgment approach to attitude change. In C. W. Sherif & M. Sherif (Eds.), *Attitude, ego-involvement, and change* (pp. 105–139). New York; London: Wiley.

Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1–83). Chicago: University of Chicago Press.

Singh, R., & Teoh, J. B. P. (2000). Impression formation from intellectual and social traits: Evidence for behavioural adaptation and cognitive processing. *British Journal of Social Psychology*, *39*, 537–554. https://doi.org/10.1348/014466600164624

Skowronski, J. J. (2002). Honesty and intelligence judgments of individuals and groups: The effects of entity-related behavior diagnosticity and implicit theories. *Social Cognition*, *20*(2), 136–169. https://doi.org/doi:10.1521/soco.20.2.136.20993

Skowronski, J. J., Betz, A. L., Thompson, C. P., & Shannon, L. (1991). Social memory in everyday life: Recall of self-events and other-events. *Journal of Personality and Social Psychology*, *60*(6), 831–843. https://doi.org/10.1037/0022-3514.60.6 .831

Skowronski, J. J., & Carlston, D. E. (1987). Social judgement and social memory: The role of cue diagnositicity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105 .1.131

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology, 22,* 435–452. https://doi.org/10.1002/ejsp.2420220503

Tausch, N., Kenworthy, J. B., & Hewstone, M. (2007). The confirmability and disconfirmability of trait concepts revisited: Does content matter? *Journal of Personality and Social Psychology*, *92*(3), 542–556. https://doi.org/10.1037/0022-3514.92.3.542

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1),

67–85. https://doi.org/10.1037/0033-2909.110.1.67

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127. https://doi.org /10.1093/scan/nsn009

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10 .1016/j.tics .2008.10.001

Trafimow, D., Bromgard, I. K., Finlay, K. A., & Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin*, *31*(7), 935–948. https://doi.org/10.1177/0146167204272179

Tribus, M. (1961). *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*. Princeton, NJ: Van Nostrand.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van der Lee, R., Ellemers, N., Scheepers, D., & Rutjens, B. T. (2017). In or out? How the perceived morality (vs. competence) of prospective group members affects acceptance and rejection. *European Journal of Social Psychology*, *47*(6), 748–762. https://doi.org/10.1002/ejsp.2269

Walker, W. R., Vogl, R. J., & Thompson, C. P. (1997). Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology*, *11*(5), 399–413. https://doi.org/10.1002/(sici)1099-0720(199710)11:5<399::aid-acp462>3.3.co;2-5

Welbourne, J. L. (1999). The impact of perceived entitivity on inconsistency resolution for groups and individuals. *Journal of Experimental Social Psychology*, *35*(5), 481–508. https://doi.org/10.1006/jesp.1999.1387

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*(3), 327–335. https://doi.org/10.1037/0022-3514.64.3.327

Wyer, R. S. (1974). *Cognitive organization and change: An information processing approach*. Potomac, MD: Erlbaum.

Ybarra, O. (2002). Naive causal understanding of valenced behaviors and its implications for social information processing. *Psychological Bulletin*, *128*(3), 421–441. https: //doi.org/10.1037 //0033-2909.128.3.421