In: B. Gawronski & G. V. Bodenhausen (Eds.). (2015). Theory and Explanation in Social Psychology.
New York: Guilford Press.

4

Social-Cognitive Theories

Bertram Gawronski
Galen V. Bodenhausen

In naming our species in his biological taxonomy, Linnaeus (1758) chose *Homo sapiens*, designating us as "the wise/knowing man." Explicit in this choice is the belief that the construction of meaningful knowledge is the preeminent characteristic separating our species from its biological cousins. In *Descent of Man*, Darwin (1871) further underscored the unique role of subjective meaning in shaping human emotion and behavior. He described, for example, the profound feelings of revulsion and palpable physical symptoms that a devout Hindu man might instantaneously feel upon discovering that he has inadvertently eaten food that is considered "unclean" for religious reasons; however, a person with a different religious belief system might eat the very same food with great relish. Examples of this sort make it clear that human behavior is profoundly influenced by our subjective understandings of the world. The notion that social reality is mentally construed and humans act and react to it on the basis of this constructed understanding forms the core of social cognition research. In general terms, social cognition research seeks to understand the mental processes through which social meaning arises and exerts its influence on behavior.

The scientific challenges inherent in studying these processes are formidable. While it may seem trivially obvious that religious beliefs, to use Darwin's example, can indeed exert a powerful effect on behavior, carefully

unpacking the mental underpinnings of such effects is far from trivial. To study such matters, psychological measures must be devised. Psychometricians have studiously devoted themselves to the creation of such tools, but psychological scientists still lack universally shared, core metrics of the sort available in the physical sciences. Perhaps the closest psychologists have come to developing a universal currency for the study of meaning is Osgood, Suci, and Tannenbaum's (1957) semantic differential approach, which empirically identified three core dimensions underlying the meaning of all concepts (i.e., evaluation, potency, and activity) and provided a standard method for measuring these dimensions. However, despite Osgood and colleagues' ambition to provide a common conceptual and operational foundation for the psychological study of meaning, psychological researchers have made relatively little use of it (for a noteworthy exception, see the work on affect control theory; e.g., Heise, 2007).

One likely reason for this neglect is the tendency for social cognition researchers to prefer to theorize in terms of general mechanistic processes through which knowledge representations influence behavior, rather than focusing on the semantic content of the representations per se (see Kashima, Chapter 3, this volume). Using Darwin's example, social cognition researchers tend to be less interested in the specifics of *what* the Hindu man believes, compared to *how* his religious beliefs influence his actions and reactions. To understand such phenomena, social psychologists have proposed a general, yet functionally distinct, class of representations (i.e., sacred values) that exert systematic effects on information processing in a manner that generalizes across a variety of different specific beliefs (e.g., Tetlock, 2003). Typically, social-cognitive theories can be said to aim for content-independence in that they try to identify general principles that apply to all sorts of belief systems. For example, although attitude research can be described as being concerned with the psychological role of one of Osgood et al.'s (1957) core dimensions of meaning, evaluation, social-cognitive research on attitudes focuses primarily on questions that are content-independent, such as: (1) How are attitudes formed? (2) How are attitudes activated? (3) How do attitudes guide behavior? (4) How are attitudes represented? (De Houwer, Gawronski, & Barnes-Holmes, 2013).[1]

Yet, as we will see, there are many challenges in constructing scientifically useful theories of this kind. Fundamentally, psychological measures cannot directly assess mental processes and representations. At best, psychometrics can only tap the overt behavioral correlates of such inner mental states (De Houwer, 2011). The thorniness of this problem led the behaviorists to reject the scientific value of mental explanations *in toto*, but few social psychologists are similarly inclined to exclude mental explanations. In this chapter, we endeavor to spell out some of the biggest theoretical challenges facing researchers who propose mental accounts for observed patterns of behavior, and we offer some potential strategies for successfully tackling them.

Social Cognition as a Level of Analysis

Since its inception in the 1970s, there have been recurring debates about the most appropriate way to conceptualize *social cognition*. Whereas some conceive of social cognition as a methodological approach to understanding social phenomena (e.g., Hamilton & Carlston, 2013), others argue that social cognition should be understood as a particular content area (e.g., Macrae & Miles, 2012). Adopting a metatheoretical view, we think that social cognition may be best characterized as a level of analysis in the study of social phenomena in that social cognition research aims at understanding social phenomena on the basis of their underlying mental processes.

The metatheoretical implications of this conceptualization can be illustrated by means of Marr's (1982) distinction of three levels of analysis: (1) the computational level, (2) the algorithmic level, and (3) the implementational level (see De Houwer & Moors, Chapter 2, this volume). According to Marr, the main goal of research at the *computational level* is to identify which types of inputs produce which kinds of outputs. In functional terms, the relevant inputs may include any type of environmental stimulus and the contextual conditions under which it is encountered, whereas outputs refer to overt behavioral responses that are elicited by a given stimulus. For example, a large body of research on behavioral priming can be described as computational in that it focuses on the particular behaviors that are elicited by exposure to various kinds of prime stimuli (for a review, see Bargh, 2006). Research of this kind differs from research at the *algorithmic level*, which is concerned with the mental mechanisms that translate inputs into outputs. This level of analysis resonates with the agenda of social-cognitive research, which aims at identifying the mental processes and representations underlying social phenomena. For example, expanding on the identification of input–output relations in studies on behavioral priming, a considerable body of research aimed at identifying the mental processes and representations that mediate the effects of prime exposure on overt behavior (e.g., Cesario, Plaks, & Higgins, 2006; Loersch & Payne, 2011; Wheeler, DeMarree, & Petty, 2007). Finally, research at the *implementational level* is concerned with the physical systems that implement the mechanisms identified at the algorithmic level. In social psychology, this approach is prominently reflected in the emerging field of social neuroscience, which aims at identifying the neural underpinnings of social phenomena (see Beer, Chapter 9, this volume). For example, expanding on mental process theories of prime-to-behavior effects (e.g., Cesario et al., 2006; Loersch & Payne, 2011; Wheeler et al., 2007), research at the implementational level may investigate the neural underpinnings of the mechanisms that mediate observed relations between prime stimuli and overt behavior.

In terms of Marr's (1982) framework, social-cognitive theories are located at the algorithmic level in that they are concerned with the mental processes and representations that mediate relations between socially rel-

evant inputs and outputs. Although this conceptualization may seem rather trivial, it helps to clarify the empirical phenomena social-cognitive theories aim to explain (*explanandum*) and the theoretical assumptions they propose to explain these phenomena (*explanans*). From an epistemological point of view, research at the computational level aims at explaining observed outputs by relating them to inputs that cause these outputs. Using the above example of behavioral priming, exposure to a particular stimulus may serve as an explanation for an observed behavioral response to the extent that the stimulus can be said to cause the behavioral response. In other words, the observed behavior represents the phenomenon that needs to be explained, and exposure to the prime stimulus serves as the event that is supposed to explain the behavior (*causal explanation*). However, stating that exposure to the prime explains the behavioral response does not say anything about *how* the prime caused the observed behavior. This question is central in research at the algorithmic level, in which the causal relation between prime exposure and behavior represents a phenomenon that is in need of further explanation (De Houwer, 2011). Research at the algorithmic level provides an answer to this question by identifying the mental mechanisms that mediate the link between prime exposure and overt behavior (*mechanistic explanation*). In this sense, social-cognitive theories offer explanations of identified input-output relations by specifying the mental mechanisms that translate inputs into outputs.[2]

Some Principles of Social-Cognitive Explanation

A central requirement for any scientific explanation is that the explanans should be conceptually independent of the explanandum (Hempel, 1970). To illustrate this requirement, imagine that Sally is wondering why Bob is not married. Telling Sally that Bob is a bachelor does not provide a useful answer to her question, because the proposed explanans (i.e., Bob is a bachelor) has conceptual overlap with the explanandum (i.e., Bob being unmarried). To qualify as a useful explanation, any answer to Sally's question should refer to something that is conceptually independent of the fact that Bob is not married (e.g., Bob's personality).

At the computational level, there is little confusion about this requirement because the concepts that are used to categorize inputs (i.e., stimuli) are rarely conflated with the concepts that are used to describe outputs (i.e., behavior). At the algorithmic level, however, the independence requirement is often violated when input–output relations are equated with the mental constructs that are proposed explain them. Using mental constructs to describe behavioral effects is problematic because it can lead to circular explanations and conceptual contradictions (De Houwer et al., 2013).

As an example, consider research showing that a short period of distraction can lead to better decisions compared to an equally long period of delib-

eration (for a meta-analysis, see Strick et al., 2011). Drawing on Dijksterhuis's (2004) original explanation, this phenomenon is often referred as the *unconscious thought effect*. From a meta-theoretical perspective, such descriptions are problematic because they equate the observed effect with the mental process that is supposed to explain the effect. That is, unconscious thought (the explanans) is empirically defined as the beneficial effect of distraction on decision quality (the explanandum). Such equations make explanations in terms of unconscious thought circular because the explanans involves the same concepts as the explanandum, thereby violating the requirement of conceptual independence. Descriptions of observed effects in terms of mental mechanisms can also lead to conceptual contradictions, for example, when differences between distraction and deliberation conditions are found to be driven by conscious overthinking in the deliberation condition (e.g., Payne, Samper, Bettman, & Luce, 2008). In this case, one would have to draw the paradoxical conclusion that the effects of unconscious thought are the product of conscious thought. Such theoretical pitfalls can be avoided by clearly distinguishing between the causal relations between inputs and outputs that need to be explained (e.g., effects of distraction on decision quality) and the mental constructs that are proposed to explain the identified input–output relations (e.g., unconscious thinking).

Although a clear conceptual distinction between mental constructs and behavioral effects is a necessary precondition for scientifically sound explanations, it is not sufficient to prevent explanatory circularity. Another potential pitfall is the lack of a clear specification of the mental mechanisms that translate inputs into outputs. This limitation can lead to circular explanations even when there is a clear conceptual distinction between the behavioral effects that need to be explained and the mental constructs that are proposed to explain them. As an example, consider the distinction between System 1 processing and System 2 processing in prominent dual-systems theories of judgment and decision making (see Deutsch, Chapter 7, this volume). Although the distinction between the two kinds of processing subsumes several conceptually distinct dualities (e.g., Kahneman, 2003), it is sometimes boiled down to the distinction between resource-dependence and resource-independence. If a given effect is attenuated by time pressure or distracter tasks, it is explained in terms of System 2 processing. Conversely, if a given effect is unaffected (or enhanced) by time pressure or distracter tasks, it is explained in terms of System 1 processing (e.g., Dhar & Gorlin, 2013). Such explanations meet the above criterion of conceptual independence between explanans and explanandum to the extent that the to-be-explained behavioral effect is described without reference to the distinction between System 1 and System 2 processing. However, they may still be circular if there is no specification of System 1 and System 2 processing over and above the assumption about their differential resource-dependence. In the absence of such specifications, claims that a given effect is due to System 1 or System 2 processing do not provide anything beyond simple classifica-

tions of observed effects (Gawronski, Sherman, & Trope, 2014). To the extent that a given effect is resource-independent it will be categorized as being due to System 1 processing, but it will be attributed to System 2 processing if it is resource-dependent. Moreover, if an effect that was initially attributed to System 1 processing turns out to depend on cognitive resources, this effect would be recategorized as the product of System 2 processing, and vice versa. Without a clear specification of the mental operations that characterize System 1 and System 2 processing, explaining resource-independent effects in terms of System 1 processing and resource-dependent effects in terms of System 2 processing involves a circular explanatory structure. Such explanations may also be criticized as irrefutable because they do not imply any prediction that could be inconsistent with a given result (see Gawronski & Bodenhausen, Chapter 1, this volume). Thus, over and above the requirement that behavioral effects should not be described in terms of the mental constructs that are proposed to explain them (i.e., conceptual independence of *explanans* and *explanandum*), social-cognitive theories should provide clear specifications of the mental mechanisms that translate inputs into outputs to avoid the criticism that they provide circular and irrefutable explanations.

How Can We Test Social-Cognitive Theories?

From a naïve point of view, one could argue that social-cognitive theories can be tested by measuring the hypothesized mental constructs and then testing whether these constructs account for the relations between inputs and outputs they are supposed to mediate. For example, drawing on the available statistical tools for testing mediation (e.g., Baron & Kenny, 1986; Preacher & Hayes, 2004), researchers may experimentally manipulate certain factors at the input level and then test whether their influence on a given output variable is statistically mediated by the measure of the hypothesized mental construct. Although this approach is rather common, it involves a number of metatheoretical problems, the most important being the fact that, as we have already noted, it is not possible to directly measure mental constructs. In a strict sense, psychological measures capture the behavioral outputs of mental processes and representations, but these outputs are conceptually distinct from their mental antecedents (De Houwer, 2011).

We already discussed the difference between mental constructs and behavioral effects when we explained the requirement that *explanans* and *explanandum* have to be conceptually independent. Yet, the impossibility of measuring mental constructs goes beyond the problem of explanatory circularity in that it also involves the measurement of mental constructs that are independent of the input–output relations they are supposed to explain. The only requirement of the independence criterion is that the input–output relations that need to be explained are conceptually distinct from the mental constructs that are proposed to explain them. To the extent that the hypoth-

esized mental construct is measured independently of the relevant input–output relations, there would be no problem with the required independence of *explanans* and *explanandum*. Yet, the possibility of measuring mental constructs is undermined by a more fundamental problem: the absence of a bi-conditional relation between mental constructs and behavioral responses (De Houwer et al., 2013). In a strict sense, direct measurement of a mental construct by means of behavioral outputs presupposes that there is a one-to-one relation between the mental construct and a particular behavior, such that variations in one unambiguously reflect variations in the other (i.e., if *p*, then *q* and, at the same time, if *q*, then *p*).

Claims of such bi-conditional relations seem untenable because (1) there can always be conditions under which the mental construct does not produce the relevant behavior and (2) the same behavior may be produced by another mental construct. For example, self-reported evaluations do not unambiguously reflect mental attitudes because the impact of mental attitudes on evaluative judgments can sometimes be disrupted (e.g., when people are motivated to conceal their attitudes; see Fazio, 2007) and evaluative judgments can be influenced by various factors other than mental attitudes (e.g., incidental mood states; see Schwarz, 1990). Similar concerns apply to the use of less reactive measures. For example, evaluative priming tasks will provide unambiguous indices of mental attitudes only if variations in mental attitudes are both necessary and sufficient to produce variations in priming effects. Yet, evaluative priming effects are influenced by various factors other than mental attitudes (e.g., processes involved in response interference; see Gawronski, Deutsch, LeBel, & Peters, 2008) and the impact of mental attitudes on evaluative priming can be reduced under certain conditions (e.g., through strategic counteraction; see Teige-Mocigemba & Klauer, 2013). To be sure, both measures can be interpreted as capturing evaluative responses in the sense of behavioral outputs. However, in the absence of a bi-conditional relation between behavioral responses and a particular mental construct, it is not possible to treat these responses as a measure of this construct. Moreover, even if there were a uni-conditional relation such that variations in a mental construct always produce variations in a given behavior, inferring the mental construct on the basis of the behavior would involve the fallacy of affirming the consequent in that the conditional "if *p*, then *q*" is used to draw the logically invalid inference "if *q*, then *p*." Such inferences are problematic, because they presuppose that the relevant behavior can never be produced by an alternative mechanism (see Gawronski & Bodenhausen, Chapter 1, this volume).

If we can measure only the outcome of mental constructs, but not mental constructs per se, how is it possible to test social-cognitive theories? Does this mean that social-cognitive theories are unfalsifiable? To answer this question, it is useful to consider our earlier discussion of Marr's (1982) levels of analysis and the mutual relations between the computational and the algorithmic levels. Specifically, we argued that social-cognitive theories at

the algorithmic level provide mechanistic explanations of causal relations between inputs and output at the computational level. From this perspective, social-cognitive theories can be tested by deriving predictions about input–output relations from their assumptions about mental processes and representations. To the extent that these predictions are empirically confirmed, researchers can treat this evidence as (preliminary) support for their theories. In this case, it makes sense to believe that the processing and representation of social information are characterized by the principles stated by the theory. However, if the predictions about input–output relations are disconfirmed, the conflict between prediction and data suggests that at least one of the assumptions that has been used to derive the prediction must be false (see Gawronski & Bodenhausen, Chapter 1, this volume). In this case, researchers are faced with the challenging task of identifying which component of this broader set of assumptions led to the conflict between prediction and data.

Although regression-based approaches to testing mediation have been criticized for a variety of reasons (e.g., Jacoby & Sassenberg, 2011; Spencer, Zanna, & Fong, 2005), the current conceptualization also clarifies why experimental approaches are better suited to test hypotheses about mental processes and representations. Because psychological measures capture only the behavioral output of mental constructs, it is not feasible to measure mental mediators for regression-based mediation analyses (e.g., Baron & Kenny, 1986; Preacher & Hayes, 2004). Instead, theoretical assumptions about mental processes and representations have to be tested by deriving predictions about causal relations between inputs and outputs. In many cases, these predictions also include hypotheses about the boundary conditions of a given behavioral effect.[3] Thus, theoretical hypotheses about *mental mediation* at the algorithmic level can be tested by deriving predictions about *contextual moderation* at the computational level (Jacoby & Sassenberg, 2011). Although this conceptualization may seem to blur the distinction between mediation and moderation, the two concepts retain their original meaning in that they refer to different levels of analysis. Whereas the term *mediation* refers to the mental mechanisms that mediate input–output relations (algorithmic level), the term *moderation* refers to contextual factors that moderate input–output relations (computational level).

Another valuable insight that can be gained from this conceptualization is the mutually supportive relation between the computational and the algorithmic level of analysis (see De Houwer, 2011). On the one hand, computational research supports algorithmic research in that causal relations between inputs and outputs at the computational level provide the empirical phenomena that algorithmic research aims to explain. On the other hand, algorithmic research supports computational research to the extent that algorithmic theorizing about mental processes and representations can lead to new discoveries of previously undetected input–output relations and their boundary conditions at the computational level. Thus, although the

two levels of analysis are distinct in the sense that they differ in terms of their *explanandum* (i.e., behavioral outputs vs. input-output relations) and in terms of their *explanans* (i.e., environmental inputs that cause outputs vs. mental processes and representations that mediate input–output relations), their relation is mutually supportive in that progress at one level advances research at the other level (and vice versa).

Explanans and Explanandum of Social-Cognitive Theories

In our discussion of Marr's (1982) levels of analysis, we argued that social-cognitive theories are located at the algorithmic level in that they aim to identify the mental mechanisms underlying social phenomena. From this perspective, the explanada of social-cognitive theories are input–output relations that can be described as "social" in some sense. For example, input–output relations may be described as socially relevant if they involve either social stimuli as inputs (e.g., effects of target characteristics on judgments in research on person perception) or social behavior as outputs (e.g., effects of nonsocial stimuli on social behavior in research on behavioral priming). Although both lines of research have made valuable contributions to our understanding of social phenomena, some critics have raised concerns that social cognition researchers tend to focus primarily on the social nature of inputs, while ignoring the relevance of their nonsocial outputs (e.g., ratings, response times) for understanding social behavior in real-world settings (e.g., Baumeister, Vohs, & Funder, 2007; Macrae & Miles, 2012). Whether or not this criticism is justified is a matter of debate. Nevertheless, it is worth noting that research on the effects of social inputs on nonsocial outputs typically has a stronger impact on other fields when corresponding effects are demonstrated for social outputs.

Although social-cognitive theories differ in terms of whether they focus on social inputs, social outputs, or both, a shared characteristic is their concern with the mental processes and representations underlying the phenomena of interest. In general, social-cognitive theories aim to provide answers to at least one of four questions: (1) How are mental representations formed? (2) How are mental representations activated? (3) How do activated mental representations guide behavior? (4) How is social information represented? Whereas the first three questions are concerned with the characteristics of *mental processes*, the fourth question is concerned with the nature of *mental structures*.

Unfortunately, testing competing explanations in terms of process versus structure can be extremely difficult (Wyer, 2007). In many cases, a finding predicted by a process account may be reinterpreted in terms of competing representational accounts, and vice versa. An illustrative example is the debate about whether dissociations between implicit and explicit measures

reflect the operation of distinct mental processes (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006) or distinct mental representations (e.g., Rydell & McConell, 2006; Wilson, Lindsey, & Schooler, 2000). Because it is not possible to directly measure either of the two, the informational value of social-cognitive theories depends on precise assumptions about how the proposed mental constructs are related to environmental inputs and behavioral outputs. For example, to provide a valuable account of social phenomena, theories focusing on the nature of mental structures should also specify how the proposed representations are activated and how activated representations guide behavior. In the absence of such assumptions, representation theories become susceptible to criticism of irrefutability because they may be able to explain any possible finding in a post-hoc fashion. Although such theories are characterized by a high level of generality in the sense that they can explain a wide range of empirical results, their predictive value is typically quite low in that it is difficult to identify which input–output relations can be expected on the basis of theory in an a priori fashion (see Gawronski & Bodenhausen, Chapter 1, this volume). Thus, to provide a valuable account of social phenomena, social-cognitive theories should address more than just one of the four questions, and ideally provide answers to all of them.

Scope and Refutability of Theoretical Claims

In addition to addressing different subsets of the four central questions, social-cognitive theories vary greatly in their scope. As we have noted in our introductory chapter (Gawronski & Bodenhausen, Chapter 1, this volume), broader theories that have widespread application are generally considered more valuable than narrow ones that account only for a relatively limited range of phenomena. At the same time, broad theories can often be so general and encompassing that they essentially become irrefutable (Quine & Ullian, 1978). The dynamics of this trade-off are quite evident in social cognition research. Some explanatory accounts focus on a particular phenomenon, invoking a delimited subset of mental constructs that have direct relevance within the given domain. For example, Jones and Davis (1965) developed a theory of correspondent inference that was concerned with a very specific question: Under what conditions do social perceivers draw dispositional inferences about actors on the basis of the actors' behavior? It is an important question, with significant applications, yet it is also a relatively narrow question. It provided a basis for clear, specific, falsifiable predictions, and indeed, a central assumption of the theory was empirically disconfirmed when it was discovered that the presence of strong situational constraints failed to adequately attenuate dispositional inferences, as the theory asserted it should (Jones & Harris, 1967). As a result, new theories of dispositional inference emerged to supersede correspondent inference theory (e.g., Gilbert, 1989; Trope, 1986). Although disconfirmed in a noteworthy respect, the

theory was ultimately quite valuable in generating explanatory progress, stimulated in large part by the interest value of its failed prediction and the previously unknown phenomenon it revealed: the correspondence bias (see Jones, 1990). This example illustrates that, counter to the common preference for theories that are consistent with larger sets of potential outcomes, theories can advance science even if their predictions have been empirically disconfirmed by stimulating novel research to understand and explain the unexpected discovery (Lakatos, 1970).

Despite the scientific value of narrow theories that focus on particular phenomena, such theories involve a considerable risk of conceptual and theoretical fragmentation. This concern is prominently reflected in the desire for ambitious theories that attempt to provide very general accounts that are applicable in all domains of social cognition. One example is the theory of reasoned action (Fishbein & Ajzen, 2010), which adapts the notions of subjective expected utility and normative beliefs about social pressure into a relatively small set of assumptions about the factors guiding intention and action (see Trafimow, Chapter 12, this volume). The theory can be, and indeed has been, applied to a very wide variety of substantive topics. Another example is Anderson's theory of information integration, described most recently in his book *Unified Social Cognition* (2008), which consists of a very small set of laws said to govern the use of information in forming judgments, choices, and intentions. In the book, Anderson argues forcefully for the necessity of developing broadly integrative theory, and he laments the state of fragmentation that results from focusing on narrow, domain-specific theories whose interest value, in his view, inevitably rises and falls faddishly.

Both the theory of reasoned action and information integration theory are concerned with the process whereby individuals use multiple informational inputs to determine a response of some kind. They are similar in both proposing a valuation process whereby discrete pieces of relevant input are translated into a common subjective evaluative metric, but they differ in their assumptions about the process governing the combination of these evaluations into an overall response. Whereas Fishbein and Ajzen's theory implies that informational inputs are translated into evaluative outputs in an additive fashion, Anderson's theory implies a weighted averaging relation. This divergence immediately suggests the possibility of a competitive test, and many such tests have indeed been conducted. For example, Anderson (1965) showed that in forming social impressions, adding two pieces of moderately positive information to two pieces of very positive information about an actor resulted in a less positive impression, compared to when just the two very positive pieces of information were provided. Such an outcome is consistent with an averaging, but not with an adding, integration function. However, other findings have contradicted the averaging mechanism (e.g., Yamagishi & Hill, 1981), and it has been shown in fact that additive and averaging functions cannot easily be distinguished on the basis of the kinds of empirical tests that have been applied toward that end (Hodges,

1973). The problem, in a nutshell, is that somewhat more complex versions of each type of integration function can easily be generated to account for any given pattern of input–output relation. In other words, the competing assumptions of the two theories are essentially irrefutable because the hypothesized information integration functions are empirically ambiguous. Without the constraints of additional predictions about boundary conditions of functional relations, it would always be possible to propose different variants of additive or averaging functions that are consistent with a given set of unpredicted findings. However, such predictions are beyond the scope of the two theories because the hypothesized integration functions are assumed to underlie *all* input–output relations under *any* condition.

The prospects for refuting unconditional claims about universal mental principles is cast further in doubt when considering the typical responses inspired by the discovery of disconfirming evidence. In the case of rational actor-type theories, one response to disconfirming evidence, common for applications of expected-utility theory within economics, is to define theoretically aberrant behavior as "irrational" and thus beyond the scope of a theory of rational choice. An alternative to dismissing inconsistent evidence as irrelevant is the possibility of revising one's assumptions about the contents of the hypothesized mediators. For example, if people behave in a manner that seems counterintuitive from the perspective of expected-utility theory, theoretical claims about the representations of value and probability are often revised to make the observed outcome consistent with basic tenets of the theory. However, such strategies make theoretical explanations in terms of expected utility circular, in that expected utility (the explanans) is merely inferred from the behavior that needs to be explained (the explanandum).

In principle, a theorist could consistently resist the rejection of a favored theory by continuously revising or adding auxiliary assumptions in order to account for each new, unanticipated finding (Lakatos, 1970). Such theoretical fine-tuning can be justified only when new empirical implications can be derived from the modifications and subjected to potential falsification; otherwise, theorists are simply indulging in the patchwork quilt fallacy (Giere, 2005). The important point is that ad-hoc modifications cannot be justified merely on the basis of the original evidence that compelled them, but only on the basis of the new predictions they offer for potential falsification in novel tests. The strategy of multiple, successive ad-hoc modifications can provide an indefinite stay of execution for a cherished theory, pointing again to the problem of irrefutability. Only as long as such modifications generate novel insights, supported by new empirical tests, can the research program be considered progressive in the Lakatosian sense (see Gawronski & Bodenhausen, Chapter 1, this volume).

It is very easy to sympathize with Anderson's (2008) assertions about the desirability of broadly integrative theory in light of the undeniably fragmented state of knowledge in social psychology. Yet the refutability dilem-

mas associated with broad, general-purpose theories appear to be substantial. An alternative strategy for striving toward greater theoretical integration lies in an approach in which theories are initially developed more modestly, within a particular domain, and then their applicability in other domains is explored. The heuristic-systematic model (Chaiken, 1980), for example, was originally proposed as a theoretical account for understanding variations in the impact of persuasive messages, but in subsequent work, its implications in a diverse range of other domains were explored and a broader model was proposed (see Chen & Chaiken, 1999).

Relation to Other Theoretical Approaches

Resonating with the quest for explanatory breadth, social cognition is often regarded as a general approach that is applicable to any content domain within social psychology (e.g., Hamilton & Carlston, 2013). Because several other approaches share this feature, an important question is how social-cognitive theories are related to other types of domain-independent theories. Does social cognition compete with other overarching approaches, or do they complement each other by accounting for different aspects of social phenomena?

In our discussion of Marr's (1982) levels of analysis, we already nodded to social neuroscience, which aims at identifying the neural underpinnings of social phenomena (see Beer, Chapter 9, this volume). In terms of Marr's framework, social neuroscience can be located at the implementational level in that it is concerned with the physical systems that implement the mechanisms identified at the algorithmic level. From a metatheoretical perspective, the relation between social cognition and social neuroscience can be described in two ways. First, social neuroscience may be conceptualized as being concerned with the neural substrates of the mental processes and representations identified by social cognition. An illustrative example of this approach is research on brain mapping, which aims at identifying the brain regions that implement specific mental operations. Second, neural responses may be regarded as a particular kind of output next to overt behavior. This conceptualization resonates with the idea that well-understood neural responses may serve as alternative measures to test hypotheses about the mental mechanisms underlying social phenomena. An important difference between the two approaches is how data at one level constrain theoretical interpretations at the other level. Whereas in the first case behavioral data and their algorithmic interpretation constrain theoretical interpretations at the neural level, in the second case relations between inputs and neural outputs constrain algorithmic theories like any other behavioral outputs. However, a major issue in designing studies of the second kind is the problem of reverse inference, which can arise when a neural output is treated as a measure of a particular mental construct (Poldrack, 2006). This issue is structur-

ally equivalent to the concern about using behavioral responses as measures of mental constructs, such that it involves the logically invalid inference "if *q*, then *p*" from the conditional "if *p*, then *q*" (see Gawronski & Bodenhausen, Chapter 1, this volume). This problem does not imply that neural data cannot be used to test social-cognitive theories. Yet, studies of this kind require particular prudence in experimental design and theoretical interpretation to avoid the fallacy of affirming the consequent (see Beer, Chapter 9, this volume).

Two other approaches that have a close relationship with social-cognitive theories are emotion theories (Manstead & Parkinson, Chapter 5, this volume) and motivation theories (Dunning, Chapter 6, this volume). Although social cognition has often been criticized for ignoring the roles of emotion and motivation, the relation between research on "hot" and "cold" processes has become much less contentious than it used to be a few decades ago (Schwarz, 2000). In fact, despite the emphasis on cognition in the term *social cognition*, many recent theories aim to integrate the unique and interactive roles of affect, cognition, and motivation (e.g., Gawronski & Cesario, 2013; Higgins, 1997; Strack & Deutsch, 2004). In recognition of this development, we generally avoided references to *cognition* in the current chapter, and instead talked about mental processes and representations, terms that were intended to subsume affective, cognitive, and motivational components. Although theories of emotion and motivation have to deal with some metatheoretical issues that are unique to these domains (see Dunning, Chapter 6, this volume; Manstead & Parkinson, Chapter 5, this volume), we would argue that the issues discussed in the current chapter are relevant regardless of whether the postulated mental constructs are affective, cognitive, or motivational. The same applies to theories that explain social behavior in terms of personality systems, to the extent that these theories characterize individual differences in terms of their affective, cognitive, and motivational underpinnings (Cervone, Caldwell, & Meyer, Chapter 8, this volume).

Another approach that is often regarded as overarching in the sense that it aims at identifying "ultimate causes" of social behavior is evolutionary theory (see Ketelaar, Chapter 11, this volume). Expanding on Marr's (1982) framework, one could argue that evolutionary theories constitute a fourth level of analysis that aims at explaining the historical processes that shaped the physical systems that implement the mental mechanisms identified at the algorithmic level (Conway & Schaller, 2002). However, when it comes to historical antecedents, evolutionary accounts often compete with cultural theories (Eom & Kim, Chapter 16, this volume), the most prominent example being theories that are concerned with historical changes in social structures (e.g., Wood & Eagly, 2012). Although the two approaches are quite different in terms of their explanans, they share the goal of identifying the historical antecedents of mental processes and representations. Yet, whereas evolutionary accounts attribute their historical retention to evolved psychological mechanisms and genes (see Johnson & Penke, Chapter 10, this volume),

cultural approaches tend to locate historical retention at the level of social groups and societies (Caporael, 2001).

A final category of theories that deserves closer attention in a chapter on social-cognitive theories are formal theories, in particular computer simulations (Fiedler & Kutzner, Chapter 17, this volume) and mathematical models (Klauer, Chapter 18, this volume). Formal theories aim to simulate or quantify the role of multiple processes in the mediation of inputs and outputs. The significance of this agenda is reflected in the principle of equifinality, which refers to cases in which different combinations of processes can produce the same behavioral outcome. For example, in research on self-regulation two people may show the same behavioral response when (1) the initial impulse and inhibitory control are weak or (2) the initial impulse and inhibitory control are strong (Sherman et al., 2008). This question plays a central role in many dual-process theories, which seek to explain social phenomena in terms of the interplay of distinct automatic and controlled processes (Deutsch, Chapter 7, this volume). Formalized theories are able to capture such complex interplays by providing computer simulations and quantitative estimates of the hypothesized processes. As such, computer simulations and mathematical modeling procedures provide valuable tools for social-cognitive theorists in specifying and testing their theories about the mental mechanisms underlying input–output relations.

Summary

The main goal of this chapter was to review the metatheoretical foundation and explanatory structure of social-cognitive theories. Drawing on Marr's (1982) conceptual framework, we argued that social cognition can be described as a level of analysis, namely, the algorithmic concern with the mental processes and representations underlying social phenomena. In this sense, social-cognitive theories can be said to provide mechanistic explanations of causal relations between environmental inputs and behavioral outputs. On the basis of this conceptualization, we identified several metatheoretical criteria for evaluating social-cognitive theories. Specifically, we argued that the explanatory and predictive value of social-cognitive theories depends on: (1) a clear conceptual distinction between the input–output relations that need to be explained and the mental constructs that are proposed to explain them; (2) a clear specification of the hypothesized mental constructs; and (3) precise assumptions about how the proposed mental constructs are related to environmental inputs and behavioral outputs. To provide valuable accounts of social phenomena, we argued that social-cognitive theories should address four central questions: (1) How are mental representations formed? (2) How are mental representations activated? (3) How do activated mental representations guide behavior? (4) How is social information represented? To the extent that social-cognitive theories include precise and refutable answers

to these questions, they can provide invaluable insights through their ability to explain and predict causal relations between environmental inputs and behavioral outputs as well as their boundary conditions. Such insights are important not only for basic research on the mental underpinnings of social phenomena, but also for applications to real-world problems that aim at changing social behavior and improving social relationships.

NOTES

1. Some social-cognitive theories focus explicitly on the semantic content of social beliefs (e.g., Cuddy, Fiske, & Glick, 2008), but even in these cases, the emphasis is typically on general information processing mechanisms that act on broad categories of meaning, which can be instantiated in terms of any of a number of more specific beliefs.
2. To avoid potential confusion, it is worth noting that the distinction between causal and mechanistic explanation goes beyond Kashima's (Chapter 3, this volume) conceptualization of causal and meaning-based explanation in that both causal and mechanistic explanation are subsumed under the term *causal explanation* in Kashima's framework.
3. In some cases, predictions about boundary conditions derived from algorithmic theories also explain why a behavioral effect may be difficult to replicate at the computational level. An illustrative example is the controversy about Bargh, Chen, and Burrows's (1996) finding that participants walked slower down the hall when they were primed with the stereotype of the elderly (see Doyen, Klein, Pichon, & Cleeremans, 2012). Drawing on an algorithmic theory of the mental mechanisms underlying behavioral priming effects, Cesario et al. (2006) argued that priming effects result from perceivers preparing themselves to interact with primed social group members. An empirically confirmed prediction of their account is that participants walk slower after "elderly" priming when they hold positive evaluations of the elderly, but they walk faster after "elderly" priming when they hold negative evaluations of the elderly (and vice versa for "youth" priming). To the extent that evaluations of the elderly are distributed evenly around a neutral value, the basic priming effect will seem impossible to replicate when its underlying mental mechanism is not taken into account.

REFERENCES

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*, 394–400.

Anderson, N. H. (2008). *Unified social cognition*. New York: Psychology Press.

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

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

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396–403.

Caporael, L. R. (2001). Evolutionary psychology: Toward a unifying theory and a hybrid science. *Annual Review of Psychology*, *52*, 607–628.

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

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford Press.

Conway, L. G., & Schaller, M. (2002). On the verifiability of evolutionary psychological theories: An analysis of the psychology of scientific persuasion. *Personality and Social Psychology Review*, *6*, 152–166.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149.

Darwin, C. R. (1871). *The descent of man, and selection in relation to sex*. London: John Murray.

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, *6*, 202–209.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*, 252–287.

Dhar, R., & Gorlin, M. (2013). A dual-system framework to understand preference construction processes in choice. *Journal of Consumer Psychology*, *23*, 528–542.

Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, *87*, 586–598.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It is all in the brain, but whose brain? *PLoS One*, *7*(1), *e29081*.

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

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

Gawronski, B. (2013). What should we expect from a dual-process theory of preference construction in choice? *Journal of Consumer Psychology*, *23*, 556–560.

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

Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review*, *17*, 187–215.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, *24*, 218–225.

Gawronski, B., Sherman, J. W., & Trope, Y. (2014). Two of what? A conceptual analysis of

dual-process theories. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 3–19). New York: Guilford Press.

Giere, R. N. (2005). *Understanding scientific reasoning*. Stamford, CT: Cengage.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). New York: Guilford Press.

Greenwald, A. G., & Nosek, B. A. (2009). Attitudinal dissociation: What does it mean? In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 65–82). Hillsdale, NJ: Erlbaum.

Hamilton, D. L., & Carlston, D. E. (2013). The emergence of social cognition. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition* (pp. 16–32). New York: Oxford University Press.

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

Hempel, C. G. (1970). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*, 1280–1300.

Hodges, B. H. (1973). Adding and averaging models for information integration. *Psychological Review*, *80*, 80–84.

Jacoby, J., & Sassenberg, K. (2011). Interactions do not only tell us when, but can also tell us how: Testing process hypotheses by interaction. *European Journal of Social Psychology*, *41*, 180–190.

Jones, E. E. (1990). *Interpersonal perception*. New York: Freeman.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions. 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*, 1–24.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). Cambridge, UK: Cambridge University Press.

Linnaeus, C. (1758). *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* (10th ed.). Holmiae: Salvius.

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

Macrae, C. N., & Miles, L. K. (2012). Revisiting the sovereignty of social cognition: Finally some action. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE handbook of social cognition* (pp. 1–11). Los Angeles: Sage.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

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

Payne, J. W., Samper, A., Bettman, J. R., & Luce, M. F. (2008). Boundary conditions on unconscious thought in decision making. *Psychological Science*, *19*, 1118–1123.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*, 717–731.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). New York: McGraw-Hill.
Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*, 995–1008.
Schwarz, N. (1990). Feelings as information: Informational functions of affective states. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior*, (Vol. 2, pp. 527–561). New York: Guilford Press.
Schwarz, N. (2000). Social judgment and attitudes: Warmer, more social, and less conscious. *European Journal of Social Psychology*, *30*, 149–176.
Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*, 314–335.
Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845–851.
Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220–247.
Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., & Nordgren, L. F. (2011). A meta-analysis on unconscious thought effects. *Social Cognition*, *29*, 738–762.
Teige-Mocigemba, S., & Klauer, K. C. (2013). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition and Emotion*, *27*, 632–657.
Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*, 320–324.
Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*, 239–257.
Wheeler, S. C., DeMarree, K. G., & Petty, R. E. (2007). Understanding the role of the self in prime-to-behavior effects: The active-self account. *Personality and Social Psychology Review*, *11*, 234–261.
Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*, 101–126.
Wood W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in human behavior. *Advances in Experimental Social Psychology*, *46*, 55–123.
Wyer, R. S., Jr. (2007). Principles of mental representation. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 285–307). New York: Guilford Press.
Yamagishi, T., & Hill, C. T. (1981). Adding versus averaging models revisited: A test of a path-analytic integration model. *Journal of Personality and Social Psychology*, *41*, 13–25.