# Why did I do that? Explaining actions activated outside of awareness

**Article** *in* Psychonomic Bulletin & Review · March 2017

DOI: 10.3758/s13423-017-1260-5



**4 authors**, including:

Ana P. Gantman
New York University
**10** PUBLICATIONS **29** CITATIONS
SEE PROFILE

Peter M. Gollwitzer
New York University
**282** PUBLICATIONS **17,283** CITATIONS
SEE PROFILE

Gabriele Oettingen
New York University
**146** PUBLICATIONS **4,225** CITATIONS
SEE PROFILE

Psychon Bull Rev
DOI 10.3758/s13423-017-1260-5

BRIEF REPORT

# Why did I do that? Explaining actions activated outside of awareness

**Ana P. Gantman**[1] · **Marieke A. Adriaanse**[2] · **Peter M. Gollwitzer**[3,4] · **Gabriele Oettingen**[3,5]



**Abstract** We review the latest research investigating how people explain their own actions when they have been activated nonconsciously. We will discuss evidence that when nonconsciously activated behavior is unexpected (e.g., norm-violating, against self -standards), negative affect arises and triggers confabulations aimed to explain the behavior. Nonconsciously activated behaviors may provide a window into everyday confabulation of (erroneous) explanations for behavior, which may also affect self-knowledge. Implications for self-concept formation and intentionality are discussed.

**Keywords** Confabulation · Explanatory vacuum · Nonconscious goal pursuit · Priming

We frequently answer questions about why we acted the way we did. "Why did you take that job?" "Why did you vote for that candidate?" In many cases, the real answers to these questions may never come to light because we have little introspective access to the mental processes that led to our choices and behaviors (Nisbett & Wilson, 1977). As a result, the explanations that people provide are (at least in part) confabulations, "based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response" (p. 231). In this article, we will review the evidence that people confabulate explanations for their own behavior. We will emphasize instances in which the behavior to be explained was triggered automatically—by incidental cues in the environment—where there is emerging evidence that confabulations can both be provoked (when an experimenter asks for an explanation) and arise spontaneously (when the automatic behavior triggers negative affect by virtue of being unexpected).

Do people really generate spontaneous confabulatory explanations for their behavior? In everyday life it is often difficult (or impossible) to assess the relationship between the origin of a given behavior and a person's explanation for performing that very same behavior. As a result, one challenge for researchers to understand these confabulated explanations for behavior is to identify contexts in which relevant causes are known. As we will review, some research directly asks people to explain their behavior, yielding evidence for provoked confabulations in both clinical and nonclinical settings. Recent research has used behaviors activated outside of awareness as a case in which researchers may further understand when and in what contexts people confabulate reasons for their behavior, not only when provoked but also spontaneously.

Historically, explanations were shown to be erroneous (i.e., confabulatory) in a clinical context. Confabulations were classified as a disorder of memory (Hirstein, 2005) and related to delusions (Turner & Coltheart, 2010). In these cases, the content of the confabulation is verifiably false—a patient might describe a distant memory as a recent event. But as we will review, this behavior is not limited to clinical samples and there may be very little observable difference between confabulation and explanation (Johansson, Hall, Sikström,

✉ Ana P. Gantman
agantman@princeton.edu

[1] Psychology Department and Woodrow Wilson School of Public and International Affairs, Princeton University, Peretsman Scully Hall, Princeton, NJ 08544, USA

[2] Utrecht University, Utrecht, The Netherlands

[3] New York University, New York, NY, USA

[4] University of Konstanz, Konstanz, Germany

[5] University of Hamburg, Hamburg, Germany

Published online: 17 March 2017

Tärning, & Lind, 2006; Nisbett & Wilson, 1977). To understand when and in what contexts people confabulate explanations for their behavior, we will first give a brief overview of the research on confabulation, with an emphasis on nonclinical instances of provoked confabulation. In research in nonclinical samples, we will see that scientists describe this behavior as confabulation because they themselves experimentally observe or manipulate causes of behavior or behavioral outcomes. Explanations that do not contain these causes of behavior are taken as confabulations. To provide evidence for confabulation in nonclinical samples, we will take the case of provoked and spontaneous confabulations aimed at explaining nonconsciously activated behavior. Finally, we will discuss implications of this work for introspection and self-knowledge as well as directions for future research.

## Confabulation

*Confabulation* is a term whose origins lie in clinical psychology. Confabulation was originally considered a disorder of memory, as patients with Korsakoff syndrome with severe amnesia would report as memories events that either did not happen or had happened much earlier in the patient's life (Hirstein, 2005). The definition of confabulation expanded to include denials of ailments, known as anosognosia, misidentification syndromes such as Capgras syndrome, and the explanations of corpus callosotomy patients describing behavior derived from linguistically inaccessible content. Confabulations can be provoked, as when a patient is asked to explain a behavior, or they may occur spontaneously (Kopelman, 1987). Critically, such confabulations are genuinely believed by patients and delivered with conviction—with no intent to deceive. Dennett (1982) writes, "It is not that they lie in the experimental situation, but that they confabulate; they make up likely sounding tales without realizing they are doing it; they fill in the gaps, guess, speculate, mistake theorizing for observing" (p. 173). The confabulatory response appears to restore a sense of agentic coherence and consistency (Cooney & Gazzaniga, 2003; Dennett, 2003).

### Provoked confabulation in nonclinical samples

As in the clinical literature, there is evidence in nonclinical samples for provoked confabulations, which are given in response to a question by an authority figure (Kopelman, 1987). There are everyday situations in which people do not have access to explanations for their behavior or do not deem causes relevant for explaining their behavior, such as the position of one choice in an array of many (Nisbett & Wilson, 1977), mimicking the behavior of another person (Tanner, Ferraro, Chartrand, Bettman, & Baaren, 2008), and choice blindness (Johansson, Hall, Sikström, & Olsson, 2005). We will review each of these examples in turn. Then we will turn to instances of both provoked and spontaneous confabulations aimed at explaining behaviors activated outside of awareness.

Nisbett and Wilson (1977) demonstrated that even when people are unaware of an experimentally manipulated cause of their behavior, they easily provide an alternative explanation (rather than saying that they do not know). For example, the authors presented participants with an array of stockings as if they were in a consumer study and asked them to select their preference. Participants overwhelmingly chose the rightmost pair despite all pairs being identical. When asked the reason for their choice, participants did not mention the position of the stockings and some even refuted this as a possibility. When asked directly about the possibility of a position effect, participants denied it and felt "either that they had misunderstood the question or were dealing with a madman" (Nisbett & Wilson, 1977, p. 244). Participants appear to easily answer questions about how they made their selection and fail to appeal to factors that systematically affect behavior. Critically, participants do not say that they do not know. Instead, an explanation is readily provided that has to meet some criterion to be deemed appropriate; in this case it seems that position is too arbitrary to be relevant to a preference.

Merely mimicking the behavior of others also does not appear to participants to be a proper explanation for preferences. In one study, participants came into the lab and were given bowls of goldfish crackers and animal crackers. They were in the room with a confederate who ate either goldfish or animal crackers. Participants mimicked the eating behavior of the confederate, such that participants with an animal-cracker-eating confederate ate more animal crackers than goldfish crackers, whereas participants with a goldfish-eating confederate ate more goldfish than animal crackers. Participants were not aware that they were mimicking the experimenter, and when asked, reported a preference for the cracker that they ate more of. In this study, mimicry mediated the relationship between eating behavior and cracker preference (Tanner et al., 2008). It seems that copying another's behavior is not an appropriate way to understand one's own behavior but one's own preferences are.

The possibility that preferences can be constructed ad hoc based on our behavior (Bem 1967) has been strikingly demonstrated by the phenomenon of choice blindness (Johansson et al., 2005). Studies on this phenomenon show that people will not only confabulate reasons for arbitrary choices but they will even generate explanations for a choice that they have not made. In the case of choice blindness, participants make choices (e.g., which of two faces is more attractive) and then are asked to verbally explain their choice. The participants do not know that on some trials a trick was employed to change their apparent decision (e.g., to the other face). Participants rarely notice that their answers have been changed (e.g., participants noticed that the faces had been swapped only 13% of

the time), and are able to discuss the reasons behind their manipulated choice with apparent ease. When discussing the real and manipulated choices of attractive faces, there were no differences between explanations of manipulated and nonmanipulated choices on dimensions including emotionality, specificity, or level of detail (Johansson et al., 2005). Follow-up research has shown that choice blindness extends to multiple domains of stimuli including political and moral attitudes (Hall, Johansson, & Strandberg, 2012). It has even been found in another sensory modality, particularly audition (Lind, 2014). Note that in the case of choice blindness, the explanation for the choice is inaccessible because the choice was never made.

To further investigate the role that confabulation may play in everyday life, we focus on a particular instance in which explanations for actions are inaccessible: behavior activated outside of awareness. In these cases, participants are sub- or supraliminally primed with a goal and subsequently can be seen to enact the relevant goal-directed behavior (e.g., Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel, 2001). We take these examples because we can identify a relevant cause for behavior as the only situational difference between two groups of participants. This way we can compare the process of explanation across groups, and see whether it is sensitive to these differences in behavior activation. In addition, we take this to be an especially informative case for everyday confabulation as we are likely exposed to prime-like stimuli in everyday life.

We presume that a number of factors can bring a person into the position of explaining an action without accessibility to its purpose. For example, the reason for initiating a behavior may be forgotten, as when one walks into the next room and forgets the reason why one has walked there in the first place (Radvansky & Copeland, 2006). For the review of evidence that follows, we focus mainly on a phenomenon called the explanatory vacuum, which occurs when participants are nonconsciously primed to act in an unexpected way (e.g., against personal standards or social norms). In these cases, participants recognize that their behavior requires an explanation, and experience negative affect which can trigger confabulation.

With that said, the existence of prime-to-behavior effects has recently come into question (e.g., Cesario, 2014; Molden, 2014; Simons, 2014), and some canonical priming experiments have failed to replicate (Harris, Coburn, Rohrer, & Pashler, 2013; Shanks et al., 2013; for a full list, see https://proveyourselfwrong.wordpress.com/2015/10/13/a-list-of-successful-and-unsuccessful-high-powered-direct-replications-of-social-psychology-findings/). There are many reasons that a replication may fail to reproduce the effects of the original study that include problems in the original study such as small sample sizes (for an in-depth analysis, see the 2012 special issue of *Perspectives in Psychological Science*) as well as fluctuations in effect size due to sampling error, and unknown contextual factors that moderate the effect. Accordingly, we take it as a given that with limited evidence of this newly emerging field of study on the explanatory vacuum, all effects would benefit from multiple highly powered direct and conceptual replications. We also take it as a given that the effects of primes, by definition, are small for both empirical reasons (see meta-analysis of priming studies, $d = .35$; 95% CI [.29, .41]; Weingarten et al., 2015) and theoretical ones (if the effects of primes were reliably large, we would be slaves to the magnetic pull of our mental associations between quotidian objects and our actions; for discussion of both methodological and theoretical factors that predict behavioral priming effects, see Payne, Brown-Iannuzzi, & Loersch, 2016). Nonetheless, we want to highlight that there is still much remaining evidence for the broader claims at stake in the priming literature, specifically that our behavior can be guided by processes that we are not aware of (see, e.g., Sheeran, Gollwitzer, & Bargh, 2013). Similarly, we think that investigating how people understand their own behavior in the context of nonconscious goal pursuit is illuminating about the broader phenomenon of confabulation and introspection in nonclinical samples.

## Behaviors activated outside of awareness

Over the past 30 years, automaticity in higher order mental processing has become a core conceptual component for understanding psychological phenomena. For example, environmental cues can serve to directly activate (i.e., prime) mental processes such as goal pursuit, habits, social behavior and decision making (Bargh, Schwader, Hailey, Dyer, & Boothby, 2012). Despite years of research investigating the many ways environmental cues may trigger behavior outside of awareness, research has only just started to investigate the question of how people (who tend to infer they are the cause of their own actions; Wegner, 2002) make sense of these nonconsciously activated behaviors.

Research has begun to investigate the psychological consequences of acting without having an accessible explanation for one's own behavior, or, in other words, of "*acting in an explanatory vacuum*" (Oettingen, Grant, Smith, Skinner, & Gollwitzer, 2006). There is evidence for both "provoked confabulation" which occurs when people are probed to explain their behavior (e.g., Bar-Anan, Wilson, & Hassin, 2010) and for spontaneous confabulation when the tendency to confabulate an explanation for their behavior may be triggered without probing (e.g., as a result of experiencing negative affect; Adriaanse, Weijers, De Ridder, De Witt Huberts, & Evers, 2014; Oettingen et al., 2006; Parks-Stamm, Oettingen, & Gollwitzer, 2010).

## Nonconscious goal pursuit

Research has also shown that desired end state representations (i.e., goals; Dijksterhuis & Aarts, 2010) can be activated outside of the awareness of the actor, and that goal-directed behaviors can be triggered and deployed outside of conscious awareness (Hassin, Uleman, & Bargh, 2005). According to auto-motive theory, goals may be activated indirectly (i.e., outside of awareness) through the repeated pairing of a given situation and its related goal; the contextual cues eventually activate the goal through the established associative link (Bargh, 1990; Bargh & Gollwitzer, 1994). This model predicts that both conscious and nonconscious activation of goals should lead to similar goal attainment rates and qualities of goal striving (Bargh et al., 2001; Chartrand & Bargh, 2002). Indeed, nonconsciously activated goals exhibit hallmarks of goal pursuit. In particular, nonconscious goals, like conscious goals, lead to goal-directed action, stay active until completed, produce persistence in the face of setbacks, and resumption after interruption (Bargh et al., 2001). Nonconscious goal triggers in the environment can include the presence of a significant other (Fitzsimons & Bargh, 2003), means that are often used to attain a goal (Shah & Kruglanski, 2003), or temptations that frequently interfere with goal pursuit (Fishbach, Friedman, & Kruglanski, 2003). The goal is activated outside of awareness, but the goal striving itself (i.e., the behaviors that serve to attain the goal) can be consciously engaged in even when the source is not available. Taken together, it appears that goals can be activated outside of awareness. We assume that these are small effects (Molden 2014; Payne et al., 2016); we advocate for more high-powered replications and greater research into the conditions under which these effects are most and least likely to occur (Molden, 2014).

In principle, individuals may be unaware of one or more aspects of the processes that underlie the direct guidance of behavior by the environment. They may be unaware of the environmental cues triggering the behavior (e.g., it may be presented below the threshold for awareness), the link between the environment and the behavior (e.g., the agent is unaware of the fact that the cue is triggering the behavior), or the outcome of that process (e.g., behavioral mimicry; Chartrand, 2005). Because the agent does not know about one or more of these aspects, they can be said to be "introspectively blank"—if asked for an explanation for the behavior, they cannot provide (a veridical) one. Critically, as we will review below, the introspective blank may be experienced as unpleasant, and can be quickly and reflexively filled by confabulation.

In the case of nonconsciously activated behavior, most often, people are aware of the outcome but unaware of either the environmental cue or its relation to the outcome (e.g., behavior, decision, preference). It is these cases of nonconsciously activated behaviors that we are concerned with in the present review. Awareness of an outcome leads people to attempt to understand why that outcome occurred, particularly if that outcome is negative or unexpected (e.g., conflicts with one's self standards or social norms) and triggers negative affect. This phenomenon is referred to as the explanatory vacuum (Oettingen et al., 2006).

## Psychological consequences of acting in an explanatory vacuum

People may exhibit either provoked or spontaneous confabulation to explain behavior activated outside of awareness. In the case of provoked confabulation, the experimenter asks participants to explain their choices or actions (Bar-Anan et al., 2010), and in the case of spontaneous confabulation, participants may experience negative affect, which serves as a trigger to confabulate an explanation for the behavior (Oettingen et al., 2006; Parks-Stamm et al., 2010; Adriaanse et al., 2014; Adriaanse et al., in press; 2016). We will discuss the evidence for each in turn.

### *Provoked confabulation*

Bar-Anan and colleagues (2010) have argued that postpriming misattribution may be a common, everyday example of misattribution or confabulation. They have shown that when male participants are primed with romantic goals, they will choose a course given by a female (vs. male) instructor, regardless of the course's actual topic. However, when asked, participants expressed that the course's topic was the primary reason for their choice. In another study, participants primed with a goal to earn money were more likely to prefer a game with pictures of American presidents as they appear on American money over another game that depicted normal pictures of the same American presidents compared to those with a neutral prime. It was only after indicating their preference for either one of these games that participants received information about the games' difficulty. Participants who received information that their game was difficult reported that they liked difficult games more than participants who later learned that their game of choice was easy. When asked, participants explained their choice using a cue that was actually provided after they had already made the choice. When asked to explain their choices, we do not know whether participants simply do not consider the role of the supraliminal primes or whether they do, but then dismiss them as unlikely or unsatisfactory causes for preferences. As of now, this is the only known set of experiments to find these effects and therefore much research is needed to reproduce and expand on them.

*Spontaneous confabulation*

There is also some evidence that nonconsciously activated behaviors can lead to spontaneous confabulation. In these cases, there is no direct request for an explanation, which triggers confabulation in the sense of someone probing for an explanation. Instead, when we find evidence for spontaneous postpriming confabulation it is preceded by negative affect, which serves as a trigger to explain the unexplained behavior. Here, we use the term negative affect broadly to describe a number of affective consequences of expectation violation that may also include dissonance, uncertainty, arousal, and others (review by Proulx, Inzlicht, & Harmon-Jones, 2012). As of now, evidence for negative affect arising from nonconsciously activated behavior comes from behavior that is mismatched with expectations for behavior in some way (e.g., it is norm violating). If the behavior is fitting in context it is unlikely (though not impossible) that the actor will search for an explanation for it. For example, in the first study to investigate the affective consequences of acting in an explanatory vacuum (Oettingen et al., 2006, Study 1), participants were given an explicitly cooperative task and a goal to compete or cooperate. The goal was either consciously set or nonconsciously activated. Here, competing served as a norm-violating behavior. Participants with a nonconscious competitive goal experienced greater negative affect than those with a conscious competitive goal as well as those with both conscious and nonconscious cooperative goals. In other words, participants with a norm-violating goal experienced greater negative affect than both participants with a nonconsciously activated norm-conforming goal or a consciously set norm-violating or norm-conforming goal. Negative affect arose as a result of behavior elicited by a nonconsciously activated, norm-violating goal.

While the affective consequences of conscious versus nonconscious goal pursuit have been found to be similar in previous work—success and failure in both conscious and nonconscious goal pursuit may lead to a positive and negative mood, respectively (Chartrand & Bargh, 2002)—when successful goal pursuit involves behavior that violates expectations (e.g., violates a norm) only nonconscious (vs. conscious) goal striving lacks an apparent explanation (Oettingen et al., 2006). When an explanation is demanded by context, recognizing this "introspective blank" elicits negative affect. We interpret this finding to mean that the inability of participants to explain their norm-violating behavior led them to experience increased negative affect. Nonconsciously activated behavior triggering negative affect in this way was recently replicated in the domains of prosocial and eating behavior (Adriaanse et al., 2014).

There are some obvious parallels between the affective consequences of acting in an explanatory vacuum and classic work on cognitive dissonance (Elliot & Devine, 1994; Festinger, 1962; Stone & Cooper, 2001), which has shown that people experience discomfort (dissonance) when an inconsistency exists between a person's behavior and a her respective attitudes. However, whereas a typical dissonance study creates a situation of *insufficient* justification for the behavior (usually a soft request by the experimenter to behave in a way that is inconsistent with one's attitudes), acting in an explanatory vacuum entails a situation of *no* justification. Accordingly, in this latter, more extreme case of no justification, discomfort is experienced (Adriaanse et al., 2014). Still, acknowledging this similarity, some (Adriaanse et al., 2014; Bar-Anan et al., 2010; Parks-Stamm et al., 2010) have proposed that—very much like the misattribution to attitudes in cognitive dissonance paradigms—another psychological consequence of nonconsciously activated behavior is a tendency to misattribute the behavior to erroneous causes. For this review, we consider this form of misattribution as a kind of confabulation.

There is also evidence that the explanatory vacuum leads to spontaneous confabulation (Parks-Stamm et al., 2010), much in the way that the dissonance literature demonstrates that counterattitudinal behaviors are spontaneously interpreted as indicative of one's own attitudes (Festinger, 1962; Lieberman, Ochsner, Gilbert, & Schacter, 2001). Parks-Stamm et al. (2010) hypothesized that the increased negative affect in the nonconscious goal condition arose specifically from the lack of explanation for the behavior. The authors found that the heightened negative affect in the nonconscious goal condition could be reduced when a plausible explanation for primed competitive behavior was made available. More specifically, participants were given a cooperative task to complete with a partner, in which acting quickly was synonymous with acting competitively. Prior to this collaborative task, participants were asked to perform a seemingly unrelated task, which half of the participants had to perform quickly and the other half accurately. Of the participants in the explanatory vacuum (e.g., enacting a nonconscious competitive goal in a cooperative task), those who engaged in the prior speed task showed less negative affect than those in the accuracy task. When primed goal-directed behaviors can be explained (i.e., by having just done a task as quickly as possible), this obviates the negative affect associated with the explanatory vacuum. However, it made no difference whether participants were asked to reflect on their performance in the cooperative task, suggesting that the prior goal was automatically taken up as an explanation for norm-violating behavior. Such spontaneous confabulation in the explanatory vacuum may be a response to negative affect elicited by nonconsciosuly activated, unexpected behavior. That said, we have only a few studies demonstrating the reflexive spontaneous confabulation of explanations for nonconsciously activated behavior, and so the possibilty of reflective confabulation has not been ruled out. The possibility of reflective confabulation, as well as greater

detail on the processes that lead to confabulation, are excellent areas for future research.

*Negative affect partially mediates confabulation*

According to Parks-Stamm and colleagues (2010) spontaneous confabulation is likely a consequence of the negative affect triggered by acting in an explanatory vacuum. People are motivated to reduce the aversive state of the introspective blank, and so confabulate an explanation for their behavior as a way to do so. In other words, negative affect triggers confabulation, which is aimed at reinterpreting the behavior as having a plausible, accessible cause.

To test this proposed sequence of events, Adriaanse and colleagues (2014) analyzed whether the tendency to confabulate was indeed mediated by negative affect. In one study, participants played a video game that primed either neutral or antisocial content. Next participants were asked to complete an ostensibly unrelated task for which they were told they would receive no credit. They were asked to help a fellow student with a tedious computer task and to stop when they felt they had sufficiently helped. Then they filled out an exit survey about the new lab space in which they had taken the study. Participants primed with antisocial behavior completed fewer help trials, experienced greater negative affect, and provided a more negative evaluation of the lab space (e.g., the chair was uncomfortable) than participants primed with neutral content. Critically, the negative affect experienced after the primed antisocial behavior mediated the relationship between priming condition and lab space evaluation (i.e., confabulation). In other words, unexpected antisocial behavior led to increased negative affect, which in turn, drove participants to confabulate an explanation for their behavior (e.g., I stopped helping because the chair was uncomfortable). These findings echo classic work distinguishing cognitive dissonance from self-perception via the role of arousal (Elliot & Devine, 1994). In summary, recognizing unexpected behavior (e.g., behavior that violates social norms or personal standards) can lead to feelings of negative affect, which triggers spontaneous confabulation. Both provoked and spontaneous confabulation appear to function to explain behavior, but we do not yet know the full extent of the similarities and differences between spontaneous and provoked confabulation. Moreover, what constitutes a satisfactory explanation for behavior is another interesting avenue for future research. For instance, why are explanations such as the one described here, in which the chair is uncomfortable, deemed an appropriate explanation for not helping, while explanations about what position stockings are placed in for stocking preferences, as described in Nisbett and Wilson's (1977) study, are not?

## Moderators of the explanatory vacuum

*The behavior demands an explanation*

Not all nonconsciously activated behavior spontaneously triggers negative affect and a subsequent need to explain the behavior. Negative affect and a tendency to confabulate appear to only arise spontaneously when the nonconscious behavior *demands* an explanation, (e.g., because it violates a norm or a consciously held personal standard; Oettingen et al., 2006; Parks-Stamm et al., 2010). The moderating role of norms or standards was demonstrated in the aforementioned study by Oettingen and colleagues (2006), who reported increased negative affect only in the nonconscious, norm-violating condition, but not in the nonconscious norm-conforming condition. This moderating role of standards on negative affect was replicated and extended, demonstrating participants' tendency to confabulate in the health domain. Participants with either high or low dieting standards completed a lexical decision task to prime them with neutral or hedonic words. Then, after engaging in a subsequent taste test, which unobtrusively measured chocolate intake, confabulation was assessed by measuring to what extent participants, after reading a text proposing that cognitively demanding tasks increase cravings for sugar, retrospectively reported that the lexical decision task that had preceded the taste test was cognitively exhausting. Evidence for a mediated moderation model was obtained, suggesting that an interaction between dieting standards and the priming condition (i.e., having high dieting standards and being primed with a hedonic orientation) led to increased confabulation regarding how cognitively exhausting (and thereby chocolate consumption justifying) the lexical decision task was. Experiencing negative affect partially mediated the interaction effect of priming and personal standards on confabulation; when the prime conflicted with personal standards, negative affect arose and subsequently triggered confabulation (Adriaanse et al., 2014, Study 2).

In contrast to the studies by Oettingen et al. (2006) and Adriaanse et al. (2014), Bar-Anan et al. (2010) have demonstrated that "provoked confabulation" (Berlyne, 1972) may follow nonconsciously activated behaviors that can hardly be considered norm violating. That is, it is only in the postpriming misattribution studies that participants were explicitly asked to explain their behavior. So, rather than norm-violation evoking the need for explaining the behavior, in this latter case it is simply the explicit request to explain one's choices or actions that leads to confabulation. We suggest that it is only in the case of spontaneous confabulation that norms or standards act as a moderator and that these constitute a nonexhaustive list of possible moderators.

*When no other explanation is available*

Participants do not experience negative affect when another plausible explanation for their behavior is provided. For example, Parks-Stamm and colleagues (2010) demonstrated that when participants completed a prior task that could explain their nonconsciously activated behavior, they no longer experienced an increase in negative affect. Specifically, the authors conceptually replicated the study by Oettingen and colleagues (2006, Study 1) with the addition of a prior, seemingly unrelated study that asked half of the participants to perform quickly and half to perform accurately. Of the participants in the explanatory vacuum, those who engaged in the prior speed task showed less negative affect than those in the accuracy task, suggesting that when primed goal-directed behaviors can be explained by a previous action (i.e., by having just done a task as quickly as possible), this may obviate the negative affect associated with the explanatory vacuum.

Critically, norm violation per se is not the cause of the increase in negative affect and the tendency to search for explanations. Negative affect did not arise when the norm violation was the result of a consciously held, norm-violating goal (Oettingen et al., 2006), as the conscious goal instructions appear to serve as a satisfactory explanation. Similarly, when participants are told about the potential influence of the prime on their behavior, they do not experience negative affect or confabulate (Adriaanse, Kroese, Weijers, Gollwitzer, & Oettingen, in press). Similar to Adriaanse et al. (2014, Study 2) participants with either high or low dieting standards were included and completed a lexical decision task to prime them with neutral or hedonic words. After engaging in the taste test, participants in a "hedonic prime-and-tell condition" (but not in the regular hedonic prime condition or in the neutral condition) were told that the lexical decision task may have affected their food intake. Results showed that participants with high dieting standards in the hedonic prime condition, but not in the hedonic prime-and-tell condition, used the explanation about cognitive exhaustion as a confabulated reason to explain their indulgent behavior. Thus, this study provides additional evidence that confabulation arises as a psychological consequence of acting in an explanatory vacuum and not of norm violation in general.

*Preference for consistency*

It is possible that some individuals are more likely to experience the explanatory vacuum than others. For example, individuals who are high in preference for consistency may be more likely to notice or to care when their behavior does not match social norms of personal standards. For example, in one study (Parks-Stamm et al., 2010, Study 3), participants played a cooperative game with a partner and were given either a conscious or a nonconscious competitive goal. For those with a nonconscious goal, it was hypothesized that a lingering feeling of guilt would make participants want to act prosocially toward their partner to compensate. To test this, a Dictator Game was played, in which they could decide how many lottery tickets to share with that partner. When norm-violating competitive behavior could not be explained by an earlier conscious goal, preference for consistency positively predicted the number of tickets shared. When the earlier conscious goal could explain the norm-violating behavior, no evidence for compensatory Dictator Game sharing was found. This suggests that the more participants expect themselves to be consistent with their past goals, the less driven they are to engage in compensatory cooperative interpersonal behavior, specifically when an earlier goal can explain their antisocial behavior.

**Consequences for self-knowledge and future behavior**

Last, it appears that confabulated reasons can spill over to affect self-knowledge. Once unexpected behavior is recognized and a reason for the behavior is confabulated, this reason may become "sticky," leading to the formation of inaccurate self-knowledge that may spill over and have downstream effects. For example, in the previously mentioned study by Bar-Anan et al. (2010), participants primed with money (vs. neutral priming) preferred a game marketed with pictures of money. After they selected the game, they were told that they chose the difficult (vs. easy) game. Participants who were told that the game was difficult (vs. easy) reported that they liked more difficult games and opted for additional information regarding "how to pursue challenges." In other words, participants used an attribute of the game revealed *after* their selection to self-ascribe traits and determine subsequent behavior (Bar-Anan et al., 2010). In another study, individuals primed with romantic goals both rated their liking for the female-taught topic significantly higher and more highly endorsed the idea that they were the kind of person who was interested in her topic (a dispositional attribution) than those with no goal prime (Bar-Anan et al., 2010).

In another experiment, participants came into the lab for a 2-day study (Adriaanse et al., 2016). On the first day, their self-reported emotional eating was measured, and they were asked to watch a neutral video while eating 20 grams of both healthy (e.g., carrots) and unhealthy (e.g., marshmallows) snacks, and a baseline measure of affect was taken. On the second day, participants were randomly assigned to a bogus feedback condition in which they were told that they either ate roughly the prescribed amount of each snack (as others in the experiment had), or that they had eaten way more than prescribed (and much more than others in the experiment). Then they were asked to retrospectively report on their affective state just before they had done the snack estimation task. Despite no differences in the negative affect reported in the

moment, participants who report that they are emotional eaters, and who were told that they ate more than the norm, retroactively described themselves as feeling more negatively prior to eating. In other words, in hindsight, they described themselves as having eaten emotionally, confabulating an explanation for why they had eaten more than the norm—when they had not even done so. Although this study employed a false feedback paradigm rather than using a nonconscious activation procedure (e.g., priming), it still created a situation where people had the experience of violating a norm without having a clear explanation of why. Participants who were confronted with norm-violating behavior attributed their own behavior to their emotions prior to eating. It is possible that this creates a self-fulfilling prophecy in which believing in emotional eating creates a readily available justification for later eating behavior that may in turn affect self-knowledge and then create a higher likelihood of emotional eating in the future.

### Implications and future directions

Research on consequences of nonconsciously activated behavior is a newly emerging topic at a time where the existence of such effects has come into question. We first and foremost advocate for large-scale direct and conceptual replications of the work reviewed here. They are small effects, but possibly uniquely illuminating for a variety of important topics, including introspection, confabulation, the experience of agency, changes in self-concept over time as well as understanding intentionality and responsibility. Confabulation has been likened to the way the brain fills in blind spots to create a unified visual field. Specifically, confabulation aims to create a unified image of conscious life without gaps in memory or agentic coherence (Hirstein, 2005; Wheatley, 2009). Given the prevalence of false memories, and the feeling of will or agency (Wheatley, 2009), it is possible that cases of story-like confabulations may be prevalent. Indeed, it has been suggested that confabulated reasons for choices may be common in everyday life (Bar-Anan et al., 2010). Future research would benefit from field studies of environmental cues—particularly as they may pertain to behavior that violates self standards—to better understand the prevalence of everyday confabulation. In addition, it is not yet known what kinds of confabulations constitute satisfactory explanations.

More research is needed to understand the role that confabulation plays in the formation of self-knowledge in healthy adults. According to self-perception theory (Bem, 1967), people make inferences about their personality and character by observing their own behavior. Given that confabulations for behavior affect decisions and self-perception directly after these are made (Bar-Anan et al., 2010), it seems possible that these confabulations carry over into behavior days or weeks later. Moreover, is it possible that these confabulations are self-enforcing and activated reflexively and repeatedly. Further research is needed to understand both the extent of the downstream effects of confabulation for behavior and changes in self-concept, as well as whether confabulations affect behavior and self-perceptions reflexively.

Finally, this research has implications for understanding intentional action and responsibility. First, we do not yet know when confabulations will spontaneously involve the environment (e.g., something about the study room) or the larger self-concept (e.g., being an emotional eater). This is an important future direction as internal versus external attributions of actions may have significant consequences for feelings of blame and responsibility for the actions as well as the likelihood that the action could induce changes in the self-concept over time (Shaver, 2012). More broadly, primed behavior is considered by researchers to be unintended like other automatic actions (e.g., habits). Yet participants experience negative affect after performing unexpected, primed behaviors and show a subsequent desire to explain that behavior. This suggests that when people enact these behaviors, they likely regard them as intentional. In other words, people are driven by the experience of negative affect to come up with an explanation for their own behavior, presumably one that renders the behavior within the realm of the agent's own predictability. If this is the case, it raises important questions about perceived responsibility for these actions. For example, if a participant acts competitively or breaks his diet due to priming, should we regard the participant as responsible for the consequences? People tend to ascribe greater blame to intentional actions than unintentional ones (Ames & Fiske, 2012), but in these cases of automatic but intentional actions, the role of intention is not so clear. Future research may benefit from attempting to understand the perceived intent and responsibility for both the actor and an observer (for the very same actions) in an explanatory vacuum paradigm.

## Conclusion

People often have little access to the higher order mental processes that give rise to preferences, choices, and behaviors (Nisbett & Wilson, 1977). To compensate for this apparent incapability, people may confabulate reasons for acting, both when directly or indirectly asked to explain their behavior, or they may confabulate spontaneously when the causes of their behavior are unknown. Spontaneous confabulation is more likely when a behavior is unexpected (e.g., violates personal standards or norms), and triggers negative affect. When no

explanation is readily available (e.g., a conscious goal, relevant prior goal), people will confabulate explanations. Taken together, this suggests that when unexpected behavior demands an explanation, it is aversive, and so people provide their own explanation even at the expense of accuracy. This phenomenon has implications for understanding the formation and updating of future behavior and self-knowledge, as well as ascriptions of responsibility and intentionality in everyday life.

**Author note** The research in this paper was supported by a grant (VENI-451-11-030) from the Netherlands Organization for Scientific Research, awarded to Marieke Adriaanse.

## References


Adriaanse, M. A., Kroese, F. M., Weijers, J., Gollwitzer, P. M., & Oettingen, G. (in press). Explaining unexplainable food choices. *European Journal of Social Psychology.*

Adriaanse, M. A., Prinsen, S., De Witt Huberts, J. C., Evers, C., & De Ridder, D. T. D. (2016). "I ate too much so I must have been sad": Emotions as a confabulated reason for overeating. *Appetite, 103,* 318–323.

Adriaanse, M. A., Weijers, J., De Ridder, D. T., De Witt Huberts, J., & Evers, C. (2014). Confabulating reasons for behaving bad: The psychological consequences of unconsciously activated behaviour that violates one's standards. *European Journal of Social Psychology, 44,* 255–266.

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24,* 1755-1762.

Bar-Anan, Y., Wilson, T. D., & Hassin, R. R. (2010). Inaccurate self-knowledge formation as a result of automatic behavior. *Journal of Experimental Social Psychology, 46,* 884–894.

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of social interactions. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition* (Vol. 2, pp. 93–130). New York, NY: Guilford Press.

Bargh, J. A., & Gollwitzer, P. M. (1994). Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior. In W. D. Spaulding (Ed.), *Integrative views of motivation, cognition, and emotion* (pp. 71–124). Lincoln, NE: University of Nebraska Press.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81,* 1014–1027.

Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L., & Boothby, E. J. (2012). Automaticity in social-cognitive processes. *Trends in Cognitive Sciences, 16,* 593–605.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183–200.

Berlyne, N. (1972). Confabulation. *The British Journal of Psychiatry, 120,* 31–39.

Chartrand, T. L. (2005). The role of conscious awareness in consumer behavior. *Journal of Consumer Psychology, 15,* 203–210.

Chartrand, T. L., & Bargh, J. A. (2002). Nonconscious motivations: Their activation, operation, and consequences. In A. Tesser, D. Stapel, & J. Wood (Eds.), *Self and motivation: Emerging psychological perspectives* (pp. 13–41). Washington, DC: American Psychological Association Press.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9,* 40–48.

Cooney, J. W., & Gazzaniga, M. S. (2003). Neurological disorders and the structure of human consciousness. *Trends in Cognitive Sciences, 7,* 161–165.

Dennett, D. C. (1982). How to study human consciousness empirically or nothing comes to mind. *Synthese*, 53, 159–180.

Dennett, D. C. (2003). The self as a responding—and responsible—artifact. *Annals of the New York Academy of Sciences, 1001,* 39–50.

Dijksterhuis, A., & Aarts, H. (2010). Goals, attention, and (un)consciousness. *Annual Review of Psychology, 61,* 467–490.

Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology, 67,* 382–394.

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Paolo Alto, CA: Stanford University Press.

Fishbach, A., Friedman, R. S., & Kruglanski, A. W. (2003). Leading us not into temptation: Momentary allurements elicit overriding goal activation. *Journal of Personality and Social Psychology, 84,* 296–309.

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

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS One, 7,* e45457.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PloS One, 8,* e72467.

Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.). (2005). *The new unconscious*. Oxford, UK: Oxford University Press.

Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310,* 116–119.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15,* 673–692.

Kopelman, M. D. (1987). Two types of confabulation. *Journal of Neurology, Neurosurgery & Psychiatry, 50,* 1482–1487.

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., & Schacter, D. L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science, 12,* 135–140.

Lind, A. (2014). *Semantic self-monitoring in speech: Using real-time speech exchange to investigate the use of auditory feedback for self-comprehension* (Vol. 158). Lund, Sweden: Lund University.

Molden, D. C. (2014). Understanding priming effects in social psychology: An overview and integration. *Understanding Priming Effects in Social Psychology, 252,* 243–249.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Oettingen, G., Grant, H., Smith, P. K., Skinner, M., & Gollwitzer, P. M. (2006). Nonconscious goal pursuit: Acting in an explanatory vacuum. *Journal of Experimental Social Psychology, 42,* 668–675.

Parks-Stamm, E. J., Oettingen, G., & Gollwitzer, P. M. (2010). Making sense of one's actions in an explanatory vacuum: The interpretation of nonconscious goal striving. *Journal of Experimental Social Psychology, 46,* 531–542.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General, 145,* 1269–1279.

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences, 16,* 285-291.

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition, 34,* 1150–1156.

Shah, J. Y., & Kruglanski, A. W. (2003). When opportunity knocks: Bottom-up priming of goals by means and its effects on self-regulation. *Journal of Personality and Social Psychology, 84,* 1109–1122.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z.,…Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLOS ONE*, *8*, e56515.

Shaver, K. (2012). The attribution of blame: Causality, responsibility, and blameworthiness. Springer Science & Business Media.

Sheeran, P., Gollwitzer, P. M., & Bargh, J. A. (2013). Nonconscious processes and health. *Health Psychology, 32,* 460–473.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76–80.

Stone, J., & Cooper, J. (2001). A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology, 37,* 228–243.

Tanner, R. J., Ferraro, R., Chartrand, T. L., Bettman, J. R., & Van Baaren, R. (2008). Of chameleons and consumption: The impact of mimicry on choice and preferences. *Journal of Consumer Research, 34,* 754–766.

Turner, M., & Coltheart, M. (2010). Confabulation and delusion: A common monitoring framework. *Cognitive Neuropsychiatry, 15,* 346–376.

Wheatley, T. (2009). Everyday confabulation. In B. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy* (pp. 205–225). Oxford, UK: Oxford University Press.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2015). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin, 142,* 472–497.