# The Integrated Theory of Emotional Behavior Follows a Radically Goal-directed Approach

**Agnes Moors**

KU Leuven, Research Group of Quantitative Psychology and Individual Differences; Centre for Social and Cultural Psychology

Ghent University, Department of Experimental-Clinical and Health Psychology

Author Note: Preparation of this article was supported by Research Programme G.0223.13N of the Research Foundation - Flanders (FWO) and the Research Fund of KU Leuven (GOA/15/003). I am grateful to the commentators for their thought-provoking comments on my target arcticle. Each of them shines light on aspects and issues that merit clarification and sharpening. I also thank Frank Baeyens and Maja Fischer for absorbing discussions.

# The Integrated Theory of Emotional Behavior Follows a Radically Goal-directed Approach

The target paper and the commentaries reveal two approaches that each propose a different mechanism as the default determinant of (emotional and other) behavior: a stimulus-driven approach and a goal-directed approach. I start by reiterating the definitions of the stimulus-driven and goal-directed mechanisms. I link the latter to the belief-desire theory of intentional action (see Dixon, 2017), and I examine whether both mechanisms should be supplemented by a third mechanism, as is suggested by Sennwald, Pool, and Sander (2017). After that, I zoom in on several ways in which the two mechanisms can interface. While some authors seem to seize these interfaces to argue for a fading of the distinction between both mechanisms, I will resist such a fading. After that, I turn to a description of the two approaches. I address the various strategies raised to salvage the stimulus-driven approach (some already mentioned in the target paper, others newly raised by the commentators), and I provide arguments for moving to a radically goal-directed approach. Next, I critically examine less radical alternatives (e.g., Eickers, Loaiza, & Prinz, 2017; Parkinson, 2017) that reject a stimulus-driven approach but see ways to salvage the existence of basic emotions, and/or the existence of the entire set of emotions. I end with a few thoughts on taxonomies of emotion theories and ways to elaborate the current one. Several more detailed issues raised by commentators are interlaced in the text.

## Two Mechanisms

The distinction between stimulus-driven and goal-directed mechanisms is borrowed from the operant learning literature, where both mechanisms are defined in terms of the content of their mediating representations (Dickinson & Balleine, 1994). A stimulus-driven mechanism is one that is mediated by an association between the representation of stimulus features and the representation of a response (an [S-R] link). The stimulus features can range from concrete, perceptual ones (e.g., brightness, loudness) to more abstract, relational ones (e.g., goal in/congruence, threat value). In other words, the word “stimulus” in “stimulus-driven” also covers the meaning of stimuli. This should settle Barrett’s (2017, p. xx) concern that the stimulus-driven mechanism is not neurologically plausible. The representation of the response, also called the action tendency, can also range from being more concrete (e.g., flight, fight) to more abstract (e.g., safety seeking, defense).

A goal-directed mechanism, by contrast, is one that is mediated by an assessment of the expected utility—or short, the utility—of action options. This assessment is based on the interaction between representations of the values and expectancies of the outcomes of the action options ([S:R-$O^v$] links). I embedded the goal-directed mechanism in an action control cycle, in which the detection of a discrepancy between a stimulus and a first goal gives the impetus for a second goal to reduce the discrepancy. This reduction can be achieved either by engaging in some action (i.e., assimilation), by choosing a different first goal (i.e., accomodation), or by biasing interpretation of the stimulus (i.e., immunization; Brandtstädter & Rothermund, 2002). If the utility of one action option is higher than that of other options (other actions, accomodation, or immunization), the corresponding action tendency (i.e., third goal) is activated. After the action tendency is manifested in behavior, the outcome of this behavior serves as the stimulus input for a new action control cycle. To illustrate, if a colleague insults you, which is discrepant with a first goal to be respected at work, then the discrepancy between both elements activates the second goal to reduce the discrepancy. This may lead to the tendency to engage in one or the other action (e.g., shouting, staring through the insulter, joking, or walking away), to accomodate (e.g., give up the goal to be respected at work)[1], or to immunize (e.g., reinterpreting the insult as a desperate cry for attention), depending on the utilities of these options.

[1] Note that accomodation can explain shifts in goal priority over time (cf. Parkinson, 2017)

The first goal in this cycle provides the value of one action outcome and corresponds to what philosophers call a desire. The expectancy corresponds to what philosophers call a belief. And the action tendency corresponds to what philosophers call an intention or goal to act. This mechanism is conform to the belief-desire theory of intentional action (Bratman, 1987), as noted by Dixon (2017; see also Heyes & Dickinson, 1990). It may be good to emphasize, however, that a very thin notion of beliefs, desires, and intentions is used here. They are representations that can have any format (e.g., verbal-like or image-like) and that are unconscious by default.

In response to Russell (2017), I wish to specify that I use the term action tendency in the sense of an inclination or directedness, not in the sense of a probabilistic trend (Johansson, 1992). One way in which a system can be directed towards a behavior is by forming a representation of it so that it can precede and cause the behavior. I concur with James' (1890) ideomotor principle that all representations of behavior have dynamic features (which is why I equate representations of behavior with action tendencies). This means that they are manifested into overt behavior unless they are counteracted by other, stronger, action tendencies or by physical boundaries. Physiological activity prepares and supports the manifestation of action tendencies in overt behavior, and no additional fiat is needed. Contrary to Eickers et al.'s (2017) interpretation, the integrated theory is not a disembodied theory and neither are the original theories that went into the integration. Bodily activity is present in most episodes, but instead of being organized around vernacular emotions, it is organized around specific actions. Running asks for activation of the limbs whether it is part of a so-called fear episode or a jogging episode.

## Three mechanisms?

The dichotomy between stimulus-driven and goal-directed (i.e., in terms of the *content of representations*) should not a priori be stacked onto other dichotomies, such as the dichotomy between rule-based computation and the activation of associations (i.e., in terms of *operations*), the dichotomy between verbal-like and image-like (i.e., in terms of the *format of representations*), or the dichotomy between automatic and nonautomatic (i.e., in terms of *operating conditions*)[2]. All these dichotomies can be seen as orthogonal to each other (e.g., see de Wit & Dickinson, 2009, for an associative acount of the goal-directed mechanism). In addition, rather than arguing that each mechanism is subserved by a separate learning system (as Sennwald et al., 2017, seem to do), I listed various routes (not all of them learning procedures) that can install the representations in both mechanisms: [S-R] links can be learned via a moderate operant conditioning procedure (proper to habits), a mere pairing of stimuli and responses, and the formation of implementation intentions, but they can also be innate; [S:R-$O^v$] links can be learned via an extensive operant conditioning procedure, via observation, or via instruction, but they can also be computed or inferred online. In sum, my model is a dual *process* model based on a single dichotomy, not a dual *systems* model in which two or more dichotomies are stacked onto each other. Importantly, the dichotomy at stake refers to two mechanisms for behavior causation, not to two learning systems. Therefore Sennwald et al.'s (2017) proposal to add a Pavlovian learning system to the habitual and goal-directed learning systems cannot be considered as an elaboration of my dual process model strictly speaking. Talk of learning *systems* suggests that learning *procedures* map neatly onto behavioral *effects* as well as onto mediating *mechanisms* for behavior causation. There are good reasons to resist such a mapping (De Houwer, 2007). Nevertheless, it is a legitimate question to ask whether a Pavlovian learning procedure installs a mechanism for behavior causation over and above the stimulus-driven and goal-directed mechanisms listed so far.

[2] Behavior caused by a goal-directed mechanism is proximally caused by an action tendency, which is a type of goal or intention. Thus, goal-directed *behavior* is nonautomatic in the sense of intentional, although it can still be automatic in the sense of fast, efficient, and unconscious. However, this does not touch on the claim that the goal-directed *mechanism* itself can be automatic in every possible sense.

A Pavlovian conditioning procedure, which consists of the pairing of a stimulus (*S1*) that does not initially elicit a response, with a stimulus (*S2*) that does already elicit a response (*R2*), typically leads to a Pavlovian conditioning effect, in which *S1* alone presentations lead to a response (*R1*) that is similar, anticipatory, opponent to, or just different from *R2*. There is no consensus about the mechanism that causes *R1*. There is fair consensus that a Pavlovian conditioning procedure installs a [S1-S2] link, but this link is not by itself sufficient to cause *R1*. Different mechanisms have been proposed for the transition from the [S1-S2] link to *R1*, depending on the nature of the relation between [S1] and [S2]. This relation can be (a) purely referential, in the sense that S1 merely reminds of S2, or (b) predictive, in the sense that S1 predicts the actual occurrence of S2, which means that the person believes that S2 will actually occur. In the *referential* case, activation is spread from *S1* to [S1] on to [S2] and any further representations. Depending on whether [S2] is itself implied in a [S2-R2] link or a [S2:R2-$O^v$→R2] link to produce *R2, R1* may also be caused by the same link. Brandon, Vogel, and Wagner (2003) posit that [S1] activates only a decayed representation of [S2], which is not capable of producing a full-blown response and sometimes even produces an opponent response. In the *predictive* case, [S1] activates the belief that S2 will occur ([S2’]), which explains why *R1* is often anticipatory (i.e., prepares for *S2*). Here again, depending on whether [S2’] is itself implied in a [S2’-R2] link or a [S2’:R2-$O^v$→R2] link to produce *R2, R1* may also be caused by the same link.

Most authors agree that *R1* is not directly caused by a goal-directed mechanism ([S1:R1-S2]) because there is no contingency between *R1* and *S2* (although there may be superstituous contingency or even real contingency in the sense that some preparatory behavior might increase the positivity of a positive S2 and decrease the negativity of a negative S2; Hearst, 1979). It is also clear that *R1* is not directly caused by a stimulus-driven mechanism ([S1-R1]) because *S1* did not elicit *R1* prior to its pairing with *S2*. Yet, Timberlake (2001) suggested that due to the pairing of *S1* with *S2, S1* comes to act as a substitute of an *S1** that in natural conditions acts as a signal of *S2* and that is itself involved in an innate [S1*-R1] link. For instance, a tone may come to substitute the smell of food that in natural conditions signals food and that is itself implied in an innate [smell-salivation] link. In sum, in all the above-described scenarios, the [S1-S2] link requires a stimulus-driven mechanism ([S2-R2] or [S1*-R1]) or a goal-directed mechanism ([$S2^{(')}$:R2-$O^{v}$] or [S1:R1-S2]) to fill in the part of behavior causation. Thus, it does not seem that Pavlovian conditioning effects require a third mechanism for behavior causation.

Sennwald et al. (2017) further discuss incentive sensitization theory (Berridge & Robinson, 2003), a theory developed to explain phenomena observed in addiction that are difficult to explain with a goal-directed mechanism. For instance, addicts relapse after having been abstinent for a long time, so that their drug use cannot be explained by a desire to avoid withdrawal symptoms. Or addicts seek drugs even if the hedonic value of the drugs is minimal or even absent. Incentive sensitization theory frames these phenomena as cases of wanting without liking: The person engages in behavior (drug intake) to obtain an outcome (intoxication), even if the outcome has no positive value. This could be formalized as *S*→[S:R-O]→*R*, in which O has lost its value v. This could then indeed be called a third mechanism for behavior causation. According to an alternative interpretation, however, the outcome still has a value, but this value is not hedonic. Otherwise put, participants can have multiple goals. Drug intake may initially satisfy multiple goals. If one of those goals is no longer satisfied by drug intake, there may still be goals left that are satisfied. It could also be that drug intake becomes a goal in itself, which is indeed satisfied when the drugs are taken in.

## Interfaces

There are various ways in which stimulus-driven and goal-directed mechanisms can interface (see Moors, Boddez, & De Houwer, in press; Kotabe & Hoffmann, 2015; Wood & Rünger, 2016). Nevertheless, it is useful to keep a distinction between both mechanisms, as well as between the two approaches that are based on them.

**Interface 1**

A first type of interface comes from the very embedding of the goal-directed mechanism in an action control cycle. This is because the first step in the action control cycle—the step in which the discrepancy between stimulus and a first goal triggers a second goal to reduce the discrepancy—can be seen as a [S-R] link couched on a very high level of abstraction. The discrepancy between stimulus and goal corresponds to the stimulus feature goal incongruent, and the goal to reduce the discrepancy can be seen as a general tendency to something about it (whether that is in the form of an overt act or of the mental acts accomodation and immunization). In the goal-directed account, however, this first step is followed by a step in which the utilities of various overt action options, accomodation, or immunization are weighed. Thus, the goal-directed account goes beyond a higly abstract stimulus-driven account.

**Interface 2**

A second type of interface is suggested by the idea that a [S:R-$O^v$] link, installed via a moderate operant conditioning procedure, can transform into a [S-R] link (called a habit) if the number of operant conditioning trials is drastically increased (and if other conditions are fulfilled, such as the absence of choice; Klossek, Yu, & Dickinson, 2011). Extensive presentation of the same outcome after the same response in the presence of the same stimulus is supposed to lead to the stamping in of the link between [S] and [R] and an evaporation of [$O^v$] (Adams, 1982).

The innate [S-R] connections put forward in evolutionary accounts can be seen as further cristallizations of what once were [S:R-$O^v$] connections to our hunter-gatherer ancestors. While these ancestors had to learn that venomous snakes lead to sickness or death and that fleeing is a good way to avoid that, today we can rely on [S-R] connections that make us flee from snakes without having to process the utility of fleeing ourselves. As [S-R] links are insensitive to changes in outcome values and expectancies, however, they may cease to produce optimal behavior when organisms enter in a new environment with a new outcome structure. Venomous snakes are still harmful in our time, but not when they are behind glass in a zoo. In the latter case, fleeing is a waste of energy and may make one look ridiculous. Thus, although innate [S-R] links may originally stem from, and hence be justified and made intelligble by, prior [S:R-$O^v$] links, it is good to keep the distinction between both links intact and to realize that they are not equally flexible and hence not equally optimal.

In response to Dixon (2017), I wish to note that the term optimal should not be understood in an olympian sense. Optimality is always bounded in the sense that people have only limited sources of information at their disposal, and that they have limited motivation and capacity to process sources of information that are at their disposal (see Bechtel & Richardson, 2010). Nevertheless, I maintain that goal-directed mechanisms are more likely to provide optimal solutions because they can take into account more sources of information (i.e., both the stimulus and the values and expectancies of outcomes) than stimulus-driven ones (i.e., only the stimulus). Recall also that goal-directed mechanisms can survive limits in capacity, especially when the goal at stake is important and motivation is high.

Furthermore, I use the term optimal in a local sense. A goal-directed mechanism delivers optimal behavior in light of the short- and long-term outcomes that the person includes in her analysis *at the time* of her decision. This behavior may or may not be ultimately optimal. A stimulus-driven mechanism, on the other hand, may provide optimal behavior in the short run, if the outcome structure of the current environment happens to fit with the outcome structure of the environment in which the [S-R] link was originally established. Again, this behavior may or may not be ultimately optimal. This goes against Weidman and Tracy's (2017) claim that the behavior that is dictated by evolution is ultimately optimal. An offended person may behave aggressively because she is driven by an innate [offense-aggression] link or because aggression has the highest utility for her at that time. But the goal-directed mechanism can also dictate a different behavior. An offended person may give a friendly moralizing speech to the offender if this action option figures in her action repertoire and if she estimates that the utility of such a speech is higher than that of aggression. In this case, the evolutionarily dictated aggressive behavior is neither immediately nor ultimately optimal.

**Interface 3**

According to a third type of interface, low-level stimulus-driven mechanisms can be embedded in a higher-level goal-directed mechanism. This could be the case in skilled actions such as typing, where the overarching goal to type remains active while the concrete movements of the fingers on the keys follow a chain of $S \rightarrow$[S-R]$\rightarrow R \rightarrow O = S \rightarrow$[S-R]$\rightarrow R$ connections. Novice typers may need a goal-directed mechanism to produce the correct finger movements at first. But as typers become more skilled, the initial [S:R-$O^v$] links may transform into [S-R] links. These [S-R] links can be recruited by the overarching goal to type which initiates the typing and remains active throughout. In a similar way, skilled social communicators may recruit low-level learned or innate [S-R] links as a means to reach their overarching communicative goals. This is consistent with Lee and Anderson's (2017) proposal that expressive behavior can be understood at an "older level of utility" for low-level physical movements and a "younger level of utility" for high-level social actions. By calling both mechanisms forms of utility, however, the authors conceal that the both types are based on different mechanisms. In sum, despite the various types of interface between stimulus-driven and goal-directed mechanisms, it is crucial to keep a distinction between both mechanisms (because they may produce different behavior in some situations) as well between the approaches based on them. It is to the approaches that I now turn.

## Two approaches

### Stimulus-driven Approach

Affect program theory and discrete and dimensional appraisal theories followa a stimulus-driven approach. Such an approach analyses the features of stimuli and ties them to action tendencies. In affect program theory, [S-R] links are innate and are neurally implemented by affect programs. Each [S-R] link is said to correspond to a particular adaptive problem (as appraised) and our hunter-gatherer ancestor's best-bet behavioral solution. In discrete and dimensional appraisal theories, [S-R] connections take the form of fixed links between appraisal patterns and action tendencies.

In affect program theory and discrete appraisal theory, different links between appraisals and action tendencies ([S-R]) are also taken as the backbone of different basic emotions, with the other components as corollary to these components (e.g., physiological responses prepare and support overt behavior, and experience is the conscious reflection of all components[3]). This is not the case for dimensional appraisal theory, which assumes that [S-R] links are manifold and do not correspond to or gravitate around basic emotions. Some discrete and dimensional appraisal theories hypothesize, for instance, that a goal-incongruent stimulus that is easy/difficult to control leads to the tendency to flee/fight, yet only the discrete version takes this [S-R] link to be characteristic of fear/anger. In sum, for affect program theory and discrete appraisal theory (but not dimensional appraisal theory), evidence in favor of their hypothesized [S-R] links is also evidence in favor of the existence of basic emotions.

[3] Some theories give a more central role to experience in that they postulate it as a mediator between appraisals and action tendencies.

When empirical research reveals that evidence for a particular stimulus-driven hypothesis is inconsistent, proponents of the stimulus-driven approach take recourse to one or more of the following strategies. A first strategy can be called refinement. Researchers refine a particular stimulus feature (e.g., control) by drawing sharp distinctions between this feature and other, related, features (e.g., power, authority, status) or they split the feature into subtypes (e.g., stable vs. situational control, prospective vs. retrospective control) and they make different predictions for each of those features or subtypes. This strategy is reminiscent of the strategy to salvage basic emotions by presenting each of them as a family with different shades.

A second strategy is to seek moderators. One type of moderators are additional stimulus features (e.g., agency, un/expectedness, legitimacy). As research continues to deliver inconsistent results, the number of moderating variables is progressively increased and ever more complex patterns of stimulus features are required to explain a reasonable chunk of the variance. This strategy is reminiscent of the strategy to salvage basic emotions by invoking mixed emotions, where additional emotions are seen as moderators. Another type of moderators are additional processes such as emotion regulation. All stimulus-driven emotion theories argue that their hypothesized [S-R] links can be diluted when individuals succesfully employ regulation strategies. Moreover, because the intensity of lab-induced emotions is typically low (because of ethical constraints), emotion regulation in the lab is often successful.

A third strategy to preserve [S-R] links is to argue that these links should be understood on a higher level of abstraction (e.g., Parkinson, 2017). When research reveals that danger not always leads to fleeing but sometimes also to fighting and freezing, the hypothesis gets rephrased as the link between danger and the tendency to seek safety of which fleeing, fighting, and freezing are concrete manifestations (Bolles, 1970). This is similar to Eickers et al.'s (2017) claim that offense does not lead to the tendency to fight, but rather to a more abstract defense tendency, which can be manifested in both fighting (covering all kinds of physical and symbolic forms of aggression) ánd withdrawal. Sznycer et al. (2017) likewise reject fixed links between adaptive problems (e.g., threat to survival) and concrete behaviors (e.g., flight), but they do argue for a fixed links between adaptive problems (e.g., threat to survival) and sets of behaviors (e.g., flight, and if that is not possible, fight). Scarantino’s (2014; in press) proposal to postulate affect programs with open-ended inputs and open-ended outputs also fits in this third strategy. The [S-R] connection is preserved on a high level of abstraction so that variety and flexibility are preserved both regarding the concrete stimuli that can trigger the [S-R] link and the concrete behavior that can result from it.

My reply to this third strategy is as follows. When stimulus-driven hypotheses are framed at a very high level of abstraction, they become virtually redundant to the first step in the action control cycle that I proposed: the step in which the discrepancy between a stimulus and a first goal leads to a second goal to reduce this discrepancy (see first type of interface above). Saying that danger leads to the tendency to seek safety comes down to saying that a discrepancy between a stimulus and a first goal (in this case safety) leads to the second goal to reduce this discrepancy. Stimulus-driven hypotheses framed at this abstract level are fairly empty, for most of the explanatory work still remains to be done. If it is true that offense leads to an abstract defensive tendency, we still need to find out which factors determine whether this abstract tendency will give rise to the concrete tendency to fight or to withdraw. This is usually the point where proponents of a stimulus-driven approach allow a goal-directed mechanism to step in to take care of planning and regulation (e.g., Scarantino, 2014, in press). Ironically then, a stimulus-driven approach that formulates its stimulus-driven mechanism on the highest level of abstraction and supplements it with a goal-directed mechanism is nearly indistinguishable from a goal-directed approach that embeds the goal-directed mechanism in an action control cycle. One difference is that the stimulus-driven approach calls the stimulus-driven mechanism “emotional” and the goal-directed mechanism “non-emotional”, whereas the goal-directed approach (see below) takes the term emotional to be a descriptive feature that is applicable to entire action control cycles.

The three strategies discussed above can explain deviations from the stimulus-driven hypotheses and in this way salvage the stimulus-driven approach, but they do not provide positive evidence for this approach. Instead of merely arguing how the stimulus-driven approach can accomodate discordance (among appraisals and action tendencies, or among other components), proponents of this approach should provide evidence for concordance, at least under those conditions under which they do predict it. For instance, they should demonstrate concordance when a person has little reason or opportunity to regulate emotional components. Demonstrating concordance requires demonstrating that there is less variety among instances of the same basic emotion subset than among instances of different basic emotion subsets. Importantly, the concordance should pertain to the core components: appraisals and action tendencies. It is not sufficient to demonstrate concordance among one of the core components and its corrolary components (e.g., among appraisals and felt appraisals, or among action tendencies and physiological responses or behavior) or to demonstrate concordance among different parts of a single component (e.g., among different parts of the experience component, as in the studies by Weidman & Tracy, 2017). Moreover, the methods for manipulation and measurement should be designed in such a way that results cannot be attributed to the mediation of stereotypic scripts, at least if the [S-R] links are assumed to be innate (see below).

## Goal-directed Approach as a Radical Alternative

The integrated theory described in the target paper is not a mere blend of dimensional appraisal theory and Russell’s (2003) constructivist theory, as Eickers et al. (2017) put it. The integrated theory goes beyond the original theories in that it shakes off certain assumptions from the original theories and incorporates new ones. Indeed, the integrated theory abandons the stimulus-driven approach followed by dimensional appraisal theory and replaces it with a radically goal-directed approach. While the stimulus-driven approach seeks to discover the *stimulus* features that determine the nature of the action tendency, a goal-directed approach proposes *action* features (such as values and expectancies) as the proximal causes of action tendencies. Here, the tendencies to flee, fight, freeze, give in, or reconcile will most often depend on the utilities of these actions in the current environment. Stimulus features can still be considered as remote causes, but only to the extent that their effect is mediated by these action features.

The stimulus-driven approach adopts a default-interventionist architecture in which the stimulus-driven mechanism is the default and the goal-directed one can intervene under special circumstances. The goal-directed approach, by contrast, adopts a parallel-competitive architecture in which both stimulus-driven and goal-directed mechanisms operate in parallell, but the goal-directed mechanism takes the upper-hand and the stimulus-driven mechanism only gets to determine behavior in special cases. It should be clear from the above that, contrary to what Eickers et al. (2017, p. xx) write, the integrated theory does not cast the goal-directed mechanism in a merely regulatory role like the stimulus-driven approach does, but rather in a leading role.

The goal-directed approach is not only more parsimonious (it proposes a limited number of explanatory factors such as values and expectancies), it also has more explanatory power. The stimulus-driven approach only yields plausible predictions in typical contexts but not in atypical contexts. It may indeed be best to flee from a threatening stimulus, when the stimulus is a tiger and you risk being eaten, but not when the stimulus is the sound of one’s train entering the station and you risk missing it (Russell, 2003). The goal-directed approach, on the other hand, yields plausible predictions for any type of stimulus in any context. This is why I expect the goal-directed approach to be more fruitful than the stimulus-driven approach. Contextual variation may certainly be the rule rather than the exception, as Barrett (2012; 2017) argues, but this variation needs to be explained. And the goal-directed mechanism seems particularly suitable for this task. If a dangerous stimulus prompts different action tendencies in different contexts (e.g., to flee in a tiger context; to approach in a train context), then this might be because different actions have different utilities in both contexts (e.g., flight has the highest utility in the tiger context; approach has the highest utility in the train context).

It should be clear from the above that, unlike what Eickers et al. (2017) suggest, the integrated theory that I propose does not give a central role to stimulus valence, not even in combination with other stimulus features or appraisals (e.g., control, expectedness, agency). I have argued that the starting point of an action control cycle is the detection of a discrepancy between a stimulus and a goal, which overlaps with the apppraisal of a stimulus as more or less goal incongruent. Goal in/congruence has often been connected to negative/positive valence by researchers, but the agent need not make this connection. The detection of a discrepancy leads to the goal to reduce this discrepancy, but this link need not be mediated by a valence assessment. The behavior chosen to achieve discrepancy reduction is not fixed by patterns of goal in/congruence or valence and additional appraisals (e.g., control, agency) but depends on the utilities of the behaviors in the current context. The *value* of potential future outcomes has a central place in this utility (in addition to the expectancies of the outcomes), but this is not the same thing as the *valence* of the stimuli present.

Eickers et al. (2017) call on the vast literature that purportedly shows emotion-specific influences on behavior (and other processes) that go beyond valence. All theories agree that goal in/congruence or valence is not sufficient to understand the variety in behavior. The question is whether the variety is best understood by invoking emotions or by invoking other causes such as appraisals, stereotypic scripts, or utility principles. Nelissen, Dijker, and Zeelenberg (2007) showed that the induction of fear leads to less prosocial behavior than the induction of guilt. Fessler, Pillsworth, and Flamson (2004) showed that the induction of anger leads to more risk taking than the induction of disgust. A straigthforward interpretation is that the emotion induction procedures in these studies indeed induced different emotions and that these in turn produced the different behaviors. Alternative interpretations are that the emotion induction procedures induced different appraisals (dimensional appraisal theory), different sterotypic scripts (constructivist theory), or different utilities of the behaviors (integrated theory), and that these are responsible for the effects. If any of these alternative interpretations is correct, explanations in terms of emotions may be approximations at best (Ortony & Turner, 1990). Emotions may point in a rather imprecise way to other factors that do the actual causal work. If so, it seems more fruitful to replace emotional explanations with explanations in terms of these other factors.

Parkinson (2017, p. XX) writes that my “focus on goal-directed practical action [...] tends to marginalize any communicative functions of emotional expression and behavior”. Russell (2017) also calls for an extension of the integrated theory that includes the component of facial expresions. Facial expressions are a form of behavior, which is why I expect them to be under the sway of the same goal-directed mechanisms as other behavior (see also Dewey, 1894; Fridlund, 1994). A person may frown instead of physically attack if frowning has a higher utility for goal satisfaction in the current situation than attacking (e.g., frowning can make another person come around without incurring the cost of physical attack). The agendas of both persons may align or conflict, which may shape the further course of the interaction. The action options that figure in a person’s repertoire and the utilities of these options are to a large extent determined by prior contextual learning. Both the narrow social context (e.g., work, home) and the broader cultural context (e.g., independent vs. interdependent cultures) co-determine the outcome structure of the environment, and in this way, co-determine the utilities of the behavioral options. Behaviors that are condemned in a culture have a high cost (in that they may lead to exclusion), and as a result, they will be chosen less frequently than behaviors that are rewarded in that culture (Cohen, 2001). Like with other behavior, the goal-directed approach leaves room for a stimulus-driven influence on facial expressions (see also the third interface, above), but this influence should not be overstated (see below).

## Less Radical Alternatives?

So far, I presented three strategies (refinement, the seeking of moderators, and abstract rephrasing) that theorists have turned to in an attempt to salvage the central role of stimulus-driven mechanisms and the basic emotions that are constituted around these mechanisms. Some theorists have turned to still other strategies to salvage basic emotions. Instead of insisting that basic emotions are characterized by dedicated mental or neural stimulus-driven mechanisms, they propose weaker criteria to individuate basic emotions (see also Ortony & Turner, 1990). Most of these strategies boil down to the grounding of basic emotions in only a single component of the emotional episode, without assuming any fixed relations with other components. For instance, one criterion ties each basic emotion to a specific appraisal ([S]), corresponding to a specific adaptive problem, life task, or socially significant event (e.g., Solomon, 2002; see Parkinson, 2017). Another criterion ties each basic emotion to a specific (more or less abstract) action tendency ([R]; e.g., Frijda & Parrott, 2011; see Parkinson, 2017). These two criteria are compatible with a goal-directed approach as long as they are not glued together (otherwise they become [S-R] links).

This move raises its own questions. Which life tasks or which action tendencies should one select and on what basis? The life tasks (e.g., dealing with danger, offense, loss, success, poisoning) and the action tendencies (safety seeking, defense, giving in, broaden and build, expel) that are typically cited as the essence of basic emotions (e.g., fear, anger, sadness, happiness, disgust) do not seem to form an exhaustive or principled list. For instance, there could be danger of poisoning, and all dangers signal a potential loss. Expulsion could be at the service of safety seeking, and so could defense. To solve this problem, theorists could come up with more principled lists in which life tasks correspond to specific fundamental goals (see McDougall, 1923), or alternatively to specific junctures between stimuli and goals (see Oatley & Johnson-Laird, 1987). The first approach is domain-specific (with domains standing for goals); the second approach is domain-general (the junctures make abstraction of the type of goal involved). Theorists could also come up with a more principled list of action tendencies. For instance, four highly abstract tendencies could be distilled from the action control cycle that I sketched: the tendencies to assimilate, accomodate, immunize, and be passive. These lists may indeed be more principled, but they cannot easily be related to existing lists of basic emotions, and in this way, may fail to deliver on their promise.

Theorists not only argue that dedicated neural or mental mechanisms are unnecessary for the grounding of basic emotions, but also that they are unnecessary for safeguarding the scientific status of the entire set of emotions. Here again, several strategies have been invoked. A first strategy is to argue that there are other criteria to demarcate the set of emotions besides mechanisms. An example is appraisal theory’s criterion that stimuli in emotional episodes are appraised as more goal relevant than stimuli in non-emotional episodes. Barrett’s (2017) remark that every action performed is goal relevant and that anything else would be metabolically frivolous does not push this criterion aside. Goal relevance is a gradual criterion. Every action may stem from a stimulus that is goal relevant to some extent, but some stimuli are more goal relevant than others, and the ones that are more goal relevant are seen as more emotional. As I argued in the target paper, however, goal relevance is only a descriptive criterion for demarcation; it goes some way in describing how laypeople rank episodes from less to more emotional. It does not qualify as a scientific criterion precisely because it does not reflect a mechanism or deep structure. Indeed, stimuli that differ in goal relevance do not require different mechanisms, neither for the extraction of goal relevance (if there is indeed a goal relevance detector, it should produce all degrees of goal relevance), neither for the translation of goal relevance in the other components (which I hypothesize to most often happen via a goal-directed mechanism).

A second strategy spelled out by Eickers et al. (2017), Parkinson (2017), and Barrett (2012) is that emotions remain real in our experience as well as in the influence they exert on our behavior and social life. If a person categorizes herself as angry, she may act in line with this emotion, thereby influencing the way in which others react to her and so on. Perhaps these authors want to suggest that even if episodes taken as instances of anger do not share a dedicated causal mechanism or deep structure, they may still share a deep consequence.

I have no doubt that self-ascriptions of emotion can influence behavior, but it remains to be seen how deep this influence really is. Second, even if evidence would show that *self-ascriptions of emotion* do have a deep influence on behavior, this is not yet evidence that *emotions* have a deep influence on behavior. Rather than arguing that emotions are real in the sense of being causally efficacious, it seems more accurate to argue that self-ascriptions of or beliefs about emotions are real in this sense. If a person believes she has supernatural powers and therefore jumps out of the window, it is her belief that makes her jump, not her supernatural powers. Studying how people come to self-ascribe emotions and how this influences their behavior may be worth our time as scientists, but it  does not make the set of emotions scientific.

The goal-directed approach proposes the following tentative answers to the two questions above. People develop emotion categories through personal experience, observational learning, instructions, and via inferences (i.e., derived relational responding, see Barnes-Holmes & Hughes, 2013). The instances in these categories share scripts (Barrett, 2006; Parkinson, 2017). Scripts may take the form of networks in which nodes stand for representations of emotional components as well as other information such as labels and norms (as suggested by network theories of emotion; e.g., Bower, 1981). Emotion scripts can influence behavior via at least two routes. According to a first, stimulus-driven route, the script is activated via one of its nodes, after which activation spreads to the other nodes, including a node that stand for an action tendency.[4] According to a second, goal-directed route, information included in the scripts influences the utility assessment of action options that are up for comparison, and in this way influences action selection. For instance, if a person frames shouting as an anger response, and the anger script contains the cultural norm that anger (like other emotions) is irrational and therefore tolerated to some extent by society, this may decrease the cost of shouting, and in this way, increase the likelihood that it is selected (Averill, 2012, see Parkinson, 2017). The person may even add an extra hysterical touch to her shouting to ensure that others do also frame her behavior as angry and not just as cruel. Thus, the frenzied feature that some theorists consider to be unique to emotional behavior may not point at a stimulus-driven origin, but may instead stem from a utility assessment, which is influenced by the shared cultural belief that emotions have this frenzied feature. I conclude that the extra layer mentioned by Parkinson (2017, p. xx) can be incorporated in the goal-directed approach, without having to assume that emotions are real in a scientific sense.

## Taxonomies of theories

The target paper represents an attempt to integrate two theories (dimensional appraisal theory and Russell's, 2003, psychological constructivist theory) against the backdrop of two other theories that they are skeptical of (affect program theory and discrete appraisal theory). The aim was not to present an encompassing taxonomy of emotion theories, but a few steps

[4] If the networks include representations of values and expectancies of action outcomes and if these play a causal role in the action tendency activated, the route qualifies as goal-directed instead of stimulus-driven.

were taken toward building such a taxonomy. Building taxonomies can be done in various ways. When only a few theories are being compared, as was the case in the target paper, it is useful to use a hierarchical tree structure with categories that branch out into subcategories. As the number of to-be-compared theories increases, however, it is often better to switch to a multi-axis structure in which various theories occupy values on various axes. A first axis is constituted by the explanandum: Some theories focus on behavior, others on emotions or emotional episodes. A second axis is constituted by the explanantia: Some theories follow a stimulus-driven approach (with stimulus-driven mechanisms as the default and goal-directed mechanisms as intervenors), others follow a goal-directed approach (with both mechanisms operating in parallel, but the goal-directed mechanism dominating behavior). A third axis is constituted by whether emotion theories hold a basic emotions view or not. The second and third axes are orthogonal: Basic emotion theories can hold a stimulus-driven approach (e.g., affect program theory and some discrete appraisal theories), postulating a mental or neural stimulus-driven mechanism as the hard liquor of each basic emotion, but they can also be non-comittal to a stimulus-driven approach (e.g., other discrete appraisal theories), taking only a single component to ground the identity of each basic emotion. Non-basic emotion theories can hold a stimulus-driven approach (e.g., dimensional appraisal theory) or a goal-directed approach (e.g., the integrated theory).

The partial taxonomy outlined so far would ideally be elaborated with other theories such as network theory (discrete version: Bower, 1981; dimensional version: Lewis, 2005) and purely descriptive appraisal theory (Clore & Ortony, 2013, mentioned by Barrett, 2017). As suggested above, network theory fits well with a stimulus-driven approach. Purely descriptive appraisal theory differs from the two mechanistic versions discussed so far in that they see appraisals as features of emotional experiences while they remain silent about the mechanisms that produce these experienes. Therefore, descriptive appraisal theory is in principle compatible with discrete and dimensional appraisal theories, but also with the goal-directed theory defended here. Other theories that would merit more thorough comparison with the theories described so far are other versions of psychological constructivist theory (Barrett, 2014; Cunningham, Stillman, & Dunfield, 2013), social constructivist theories (Averill, 2012; Boiger & Mesquita, 2012; Parkinson, 1995), neoJamesian theories (e.g., Prinz, 2004; Damasio, 1999), and other belief-desire theories (Reisenzein, 2009). This is on my to-do list.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology: B, 34*(2), 77-98.

Averill, J. R. (2012). Anger and aggression: An essay on emotion. New York: Springer-Verlag.

Barrett, L. F. (2012). Emotions are real. *Emotion, 12*(3), 413-429.

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Review, 6*(4), 292-297.

Barrett, L. F. (2017). Categories and their role in the science of emotion. *Psychological Inquiry*, xx, xxx-xxx.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences, 26*(9), 507-513.

Boiger, M., & Mesquita, B. (2012). The construction of emotion in interactions, relationships, and cultures. *Emotion Review, 4*(3), 221-229.

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological Review, 77*(1), 32-48.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36,* 129-148.

Brandon, S. E., Vogel, E. H., & Wagner, A. R. (2003). Stimulus representation in SOP: I: Theoretical rationalization and some implications. *Behavioural Processes, 62*(1), 5-25.

Brandtstädter, J., & Rothermund, K. (2002). The life-course dynamics of goal pursuit and goal adjustment: A two-process framework. *Developmental Review, 22,* 117-150.

Bratman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Clore, G. L., & Ortony, A. (2013). Psychological construction in the OCC model of emotion. *Emotion Review*, *5*(4), 335-343.

Cohen, D. (2001). Cultural variation: Considerations and implications. *Psychological Bulletin*, *127*(4), 451-471.

Cunningham, W. A., Dunfield, K. A., & Stillman, P. E. (2013). Emotional states from affective dynamics. *Emotion Review*, *5*(4), 344-355.

Damasio, A. R. (1999). The feeling of what happens: Body and emotion in the making of consciousness. New York: Harcourt Brace.

Deci, E. L., & Ryan, R. M. (1980). Self-determination theory: When mind mediates behavior. *Journal of Mind and Behavior,* 1(1), 33-43.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish journal of psychology*, *10*, 230-241.

Dewey, J. (1894). The theory of emotion: I. Emotional attitudes. *Psychological Review, 1,* 553-569.

de Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: a case for animal–human translational models. *Psychological Research PRPF*, *73*(4), 463-476.

Dickinson, A. (2016). Instrumental conditioning revisited: Updating dual-process theory. In J. B. Trobalon & V. D. Chamizo (Eds.), Associative learning and cognition: Homage to Professor NJ Mackintosh. In Memoriam (1935-2015), 177-195.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1-18.

Dixon, T. (2017). Labels, rationality, and the chemistry of the mind: Moors in historical context. *Psychological Inquiry, xx*, xxx-xxx.

Eickers, G., Loaiza, J. R., & Prinz, J. J. (2017). Embodiment, context-sensitivity, and discrete emotions: A response to Moors. *Psychological Inquiry, xx*, xxx-xxx.

Fessler, D. M., Pillsworth, E. G., & Flamson, T. J. (2004). Angry men and disgusted women: An evolutionary approach to the influence of emotions on risk taking. *Organizational Behavior and Human Decision Processes*, *95*(1), 107-123.

Frijda, N. H., & Parrott, W. G. (2011). Basic emotions or Ur-emotions? *Emotion Review, 3*, 406-415.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, CA: Academic.

Hearst, E. (1979). Classical conditioning as the formation of interstimulus associations: Stimulus substitution, parasitic reinforcement, and autoshaping. In A. Dickinson & R. A. Boakes, *Mechanisms of learning and motivation: A memorial volume to Jerzy Konorski* (pp. 19-52). Hillsdale, NJ: Erlbaum.

Heyes, C., & Dickinson, A. (1990). The intentionality of animal action. *Mind & Language*, *5*(1), 87-103.

James, W. (1890). *Principles of psychology*. London: MacMillan.

Johansson, I. (1992). Intentionality and tendency: How to make Aristotle up-to-date. In K. Mulligan (Ed.), *Language, truth and ontology* (pp. 180-192). Dordrecht, The Netherlands: Springer.

Klossek, U. M., Yu, S., & Dickinson, A. (2011). Choice and goal-directed behavior in preschool children. *Learning & behavior*, *39*(4), 350-357.

Kotabe, H. P., & Hoffmann, W. (2015). On integrating the components of self-control. *Perspectives on Psychological Science*, *10,* 618-638.

Lee, D. H., & Anderson, A. K. (2017). Emotions as information processing functions in behavior and experience. Psychological Inquiry, xx, xxx-xxx.

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic system modeling.*Behavioral and Brain Sciences, 28*, 169-194.

McDougall, W. (1923). An introduction to social psychology. Boston: Luce.

Moors, A., Boddez, Y., & De Houwer, J. (in press). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*.

Nelissen, R. M., & Dijker, A. J. (2007). How to turn a hawk into a dove and vice versa: Interactions between emotions and goals in a give-some dilemma game. *Journal of Experimental Social Psychology*, *43*(2), 280-286.

Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and emotion*, *1*(1), 29-50.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review, 97,* 315-331.

Parkinson, B. (1995). *Ideas and realities of emotion*. London: Routledge.

Parkinson, B. (2017). Emotion components and social relations. *Psychological Inquiry, xx,* xxx-xxx.

Prinz, J. J. (2004). *Gut Reactions: A perceptual theory of emotion*. Oxford, UK: Oxford University Press.

Reisenzein, R. (2009). Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research*, *10*(1), 6-20.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110,* 145-172.

Russell, J. A. (2017). Impressive new theory and theorist. *Psychological Inquiry, xx,* xxx-xxx.

Scarantino, A. (2014). The motivational theory of emotions. In D. Jacobson & J. D'Arms (Eds.), *Moral psychology and human agency* (pp. 156-185). Cambridge, UK: Cambridge University Press.

Scarantino, A. (in press). Do emotions cause actions, and if so how? *Emotion Review*.

Sennwald, V., Pool, E., & Sander, D. (2017). Considering the influence of the Pavlovian system on behavior: Appraisal and value representation. *Psychological Inquiry, xx, xxx-xxx.*

Solomon, R. C. (2002). Back to basics: On the very idea of “basic emotions”. *Journal for the theory of social behaviour*, *32*(2), 115-144.

Sznycer, D., Cosmides, L., & Tooby, J. (2017). Adaptationism carves emotions at their functional joints. *Psychological Inquiry*, *xx*, xxx-xxx.

Timberlake, W. (2001). Motivational modes in behavior systems. *Handbook of contemporary learning theories*, 155-209.

Weidman, A. C., & Tracy, J. L. (2017). How to study the structure of emotions? A welcome call to action and a pragmatic proposal. *Psychological Inquiry, xx*, xxx-xxx.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, *67,* 289-314.