Erkenntnis (2022) 87:393–423
https://doi.org/10.1007/s10670-019-00200-8

ORIGINAL RESEARCH

Emotional Actions Without Goals

Isaac Wiegman[1]

Received: 20 September 2018 / Accepted: 7 December 2019 / Published online: 7 January 2020

Abstract
Recent accounts of emotional action intend to explain such actions without reference to goals. Nevertheless, these accounts fail to specify the difference between goals and other kinds of motivational states. I offer two remedies. First, I develop an account of goals based on Michael Smith's arguments for the Humean theory of motivation. On this account, a goal is a unified representation that determines behavior selection criteria and satisfaction conditions for an action. This opens the possibility that mental processes could influence behavior without such a unified representation and hence, without goals. Second, I develop a model of emotions and appetites on which behavior selection criteria can be decoupled from satisfaction conditions. If this model is correct, then in many cases, there is no unified representation that determines the behavior selection criteria and satisfaction conditions of emotional actions. In contrast with many traditional accounts of action, the model suggests the following: Whether or not a behavior constitutes an action does not depend on the agent's grasp of the behavior's aim. Rather, a behavior constitutes an action if it was organized by the agent, where an agent can organize actions without a coherent grasp of their aim. Some emotional actions are manifestations of this possibility.

1 Introduction

Even the most level-headed adults can, under the influence of emotions, act in ways they later regret. Perhaps you become more demanding of your children one week, only to realize later that you were acting out of anxiety about an uncertain prospect (e.g., a job opening, a tenure decision). Or perhaps you become uncharacteristically angry at a colleague, only to realize later that your anger originated with a disrespectful email from a student. In the moment, you may not realize the full extent of the emotion's influence. At times, it can even remain hidden until much later when

✉ Isaac Wiegman
isaac.wiegman@txstate.edu

[1] Department of Philosophy, Texas State University, 601 University Drive, San Marcos, TX 78666, USA

the perspective of a therapist or a close other reveals the nature of the emotion's influence. These phenomena (among others) give emotional actions an interesting role in philosophical theories of action. The influence of emotions is more interesting than reflexive behaviors; emotional actions appear to be organized by the agent toward some end, yet in many cases, the end eludes the agent's awareness.

Considerations such as these have led some to controversial conclusions. In a seminal paper, Rosalind Hursthouse (1991) argues that emotional actions are arational; that in the grip of emotion, the agent may have no belief that identifies the agent's reason for action, or the "favorable light" in which he viewed his action. More recently, others have extended Hursthouse's insights (e.g., Döring 2003; Kovach and DeLancey 2005). They argue that some emotional action have no goals, and they attempt to buttress these conclusions with claims about the nature of emotions and their influence on action.

On my view, these accounts fail to specify the difference between goals and other kinds of motivational states. Even if they accurately depict the nature of emotions, the status of emotional actions remains unresolved; one can reasonably doubt that any emotional actions lack goals (or so I argue in the following section). I offer two remedies. First, I develop an account of goals based on Michael Smith's arguments for the Humean theory of motivation (in Sect. 3). On this account, a goal is a unified representation that determines behavior selection criteria and satisfaction conditions. This opens the possibility that mental processes could influence behavior without such a unified representation and hence, without goals. Second, I develop a model of emotions and appetites on which behavior selection criteria can be decoupled from satisfaction conditions (in Sect. 4). If this model is correct, then in many cases, there is no unified representation that determines the behavior selection criteria and satisfaction conditions of emotional or appetitive actions. Moreover, this model offers a much more plausible explanation of emotional actions than goal-based accounts of action (or so I argue in Sect. 5). According to these accounts, actions just are behaviors that are explained by one of the agent's goal or equivalently, by the agent's desire to bring about some end (e.g., Smith 1987, 1998). In contrast with such accounts, the model on offer here suggests the following (as I point out in the conclusion): Whether or not a behavior constitutes an action does not depend on the agent's grasp of the behavior's aim. Rather, a behavior constitutes an action if it was organized by the agent, where an agent can organize actions without having a coherent grasp of their aim. Some emotional actions are manifestations of this possibility.

2 Emotional Action

As Hursthouse (1991) understands them, typical examples of emotional actions include mussing up the hair of one's child (in the grip affection or joy) or gouging out the eyes in a portrait (in the grip of hatred) or kicking a door that refuses to shut

(in the grip of anger) or rolling around in the clothing of one's deceased spouse (in the grip of grief).[1]

Hursthouse argues that actions such as these are "arational." By this, she means that they meet three descriptive conditions. First, each is done intentionally. Second, the agent does not do it for a reason. By this, Hursthouse means that there is no "true description of action of the form 'X did it (in order) to …' or 'X was trying to…' which will 'reveal the favorable light in which the agent saw what he did,'" (p. 59), and she makes clear that this means there is no belief attribution that would correctly identify the agent's reason for action. Third, the agent would not have so acted were she not in the grip of an emotion.[2]

By contrast with Hursthouse, the focus here is on the lack of goal attributions rather than belief attributions. Nevertheless, for any behavior that satisfies her description of arational action, that behavior will also lack a goal. The most natural understanding of a goal is just an outcome that an agent wants to bring about (see, e.g., Smith 1987). Moreover, for any behavior, if there were some outcome that the agent wanted to bring about and for which the agent selected that behavior, *then there would be* a "true description of action of the form 'X did it (in order) to…' or 'X was trying to…' which [would] 'reveal the favorable light in which the agent saw what he did…'" Therefore, arational actions, if any there are, are not goal-directed.

To make a preliminary case that some emotional actions lack goals, I shall consider a few examples of angry actions. First, consider cases in which someone stubs their toe on a piece of furniture or bashes their thumb with a hammer or jams their finger in a closing door. In such cases, it is not uncommon for someone to get angry and, out of anger, to kick the furniture, throw down the hammer, or punch the door. What outcome are these behaviors selected to bring about? It seems unlikely that people are attempting to bring about an outcome in which a door is punched (etc.). As Hursthouse recognizes, some of these cases will be ones in which people want to express their anger, and in those cases, the desired outcome is that anger be

[1] Hursthouse does not give a definition of emotions or emotional action, aside from saying that they are actions people take because they are in the grip of an emotion. I will follow her in this. A definition is superfluous so long as there is an identifiable phenomenon or set of examples of actions (such as those listed prior to this note), which all agree are emotional actions.

[2] Several clarifications are in order here. First, if one is committed to a dependency relation between reasons for action and goal representations, as is Smith (1987), then Hursthouse's second condition (the agent did not act for a reason) would mean that emotional actions do not have goals. But second, I think it is more plausible to say that the agent did not have a goal in acting than it is to say that she had no reason for acting. As Davidson notes, it seems natural in some cases to say that an agent's only reason for acting is that she wanted to (Davidson 2001, p. 6). Applied to cases of emotional action, we could say that sometimes the agent's only reason for acting was that she felt angry or that the action felt appropriate in some sense. I make no attempt here to argue for or against the plausibility of this claim. Rather, my aim is to show that some emotional actions do not have goals irrespective of whether they are done for reasons. If this argument is successful, then it is open to me either to agree with Hursthouse that these actions are not done for a reason or to take on Davidson's (tentative) suggestion that they are done for a reason (where the emotion is part of the reason), while denying that the agent's reason is constituted by one of her goals.

expressed. Nevertheless, it is not always plausible to attribute a desire like this when someone (a child, for instance) lashes out in anger.[3]

A second example can be found in one of James Joyce's short stories in *Dubliners*. Leading up to the event, we find Mr. Farrington in a sorry state:

> A very sullen-faced man…full of smouldering anger and revengefulness. He felt humiliated and discontented; he did not even feel drunk; and he had only twopence in his pocket. He cursed everything. He had done for himself in the office, pawned his watch, spent all his money; and he had not even got drunk… He had lost his reputation as a strong man, having been defeated twice by a mere boy. His heart swelled with fury and, when he thought of the woman in the big hat who had brushed against him and said Pardon! his fury nearly choked him. (Joyce 2011, p. 81)

Arriving home, Farrington finds "the kitchen empty and the kitchen fire nearly out" and his wife at the chapel. When one of his five children comes down to kindle the fire and make him dinner, Farrington beats him mercilessly with his walking stick, exclaiming "I'll teach you to let the fire out!" (82).

This description is a fictional example of a real phenomenon observed and studied by social psychologists: triggered displaced aggression (Marcus-Newhall et al. 2000). To study this phenomenon, psychologists typically provoke half of the participants (Pedersen et al. 2000). For example, were you in this group of participants, you would be asked to complete difficult anagrams and report your answers aloud with jarring music playing in the background. At several points in the task, a frustrated sounding voice would interrupt you and ask you to speak louder. By contrast, were you in the no-provocation group, the anagrams would be easy, the music soft, and the experimenter's interruptions neutral reports of your progress in the task. After this initial treatment, half the participants are triggered by a research assistant (distinct from the provoking experimenter): the assistant makes some reading mistakes while giving prompts and politely reports the participant's below average performance on the task. The other half of the participants are not triggered: the assistant makes no reading mistakes and reports average performance. Participants then evaluate performance of the assistant under the impression that their evaluation will affect the assistant's chances of getting a summer internship.

So, in a typical experiment, there are four groups of participants: provoked and triggered, provoked and not triggered, not provoked and triggered, not provoked and not triggered. The phenomenon of triggered displaced aggression refers to fact that only the participants who are provoked *and* triggered gave a negative evaluation of the assistant. The negative evaluation is standardly considered an act of aggression, because participants know that the evaluation will harm the assistant's prospects. By contrast, among participants who were not angry, evaluations were almost identical whether triggered or not.

This experiment seems to capture actions that are motivated by anger, but which are difficult to explain in terms of outcomes that the participants might want to

[3] For further discussion, see Hursthouse (1991), pp. 60–61

achieve (via their evaluations). It does seem as if the triggered and provoked participants are disposed to harm the assistant, but by analogy with kicking the furniture, it is not obvious that the participants desire this outcome. On what basis would be attribute such a desire? The causal connection with the prior provocation of the experimenter and the triggering circumstances vitiate many potential explanations. It cannot be that they wish to censure the assistant for politely reporting below average performance or making reading mistakes. After all, participants who were triggered without being first provoked did not have any such desire.

A more plausible suggestion is that the experimenter's provocation made the participant feel small and that the aim of aggression is to adjust their relative social status with respect to the target of aggression (cf. Nussbaum 2016, Chapter 2; Sell 2011). However, this does not explain the absence of aggression in participants who were provoked but not triggered. Even if this explanation is plausible enough for this instance of triggered displaced aggression it is unlikely to fit all the manifestations of this phenomenon.[4] Depending on culture and context, displaced aggression can easily backfire, and in ways that are all too obvious. Specifically, the agent of redirected anger often knows that the most effective course to adjusting one's status may be to respond in a cooler and more calculated manner. So, if displaced aggression were always driven by a goal of adjusting one's status, then we would not expect people often to take actions that they know will make the world *diverge* from the desired outcome. Yet, people often do act in exactly that way. Therefore, it seems unlikely that these acts of aggression are explained by a desire for the proposed outcome (adjustment of relative status).

These examples are not meant to be conclusive. Rather, their point is to explain why emotional actions appear to lack an obvious goal. Of course, many lines of argument have been explored in response to examples of this kind. For my purposes, these responses fall into two broad categories. One category of response suggests that emotions instate goals of various kinds, which explain the emotional actions under consideration and make them consistent with belief-desire explanations (which are taken to imply goals). Another category attempts to bolster Hursthouse's arguments by defending substantive claims about the nature of emotions; claims intended to show how emotions can motivate action without instating goals. For an example of the first category of response, Michael Smith has argued that emotions lead to desires to perform specific kinds of actions, such as desires to kick a door, glare or yell, etc. (Smith 1998, pp. 21–23). In a slightly different vein, Goldie (2000) has argued that these actions are explained by wishes, which include something like an imagined goal. I will address these *goal-based accounts*, and Smith's account in particular, in Sect. 5. For now, I will only suggest that these responses exploit a key difficulty in proving that any given action lacks a goal: The conceptual apparatus of goals and ends is flexible enough to extend, without apparent strain, to explain the behavior of thermostats and other mindless machines (Dennett 1997, pp. 66–67). Absent a substantive account of goals to police the appropriate boundaries, it is difficult to prove that any given action had no goal and so also proportionally

[4] As Hursthouse (1991, p. 60) points out, piecemeal explanations will not suffice.

easy to give compelling goal-based explanation. Thus, to give stronger proof that emotional actions do not have goals at all requires saying more about how emotions cause action and how this differs from the way that goals cause actions. Up until the present, notable attempts to bolster Hursthouse's arguments have attempted this and failed.

For example, Döring (2003) has argued that some emotional actions do not have ends and are instead explained by the "…emotion's affect which gives it motivational force" (p. 224). First, she rejects belief-desire explanations of emotional actions. Following Smith (1987), she points out that desires have the opposite direction of fit from beliefs: whereas desires dispose the agent to change the world to fit the content of the desire, beliefs dispose the agent to change the content of the belief to fit the world. She then suggests a critical difference between desires and emotions: "Emotions certainly do not imply that the world has to be changed in [specific ways]. An emotion need not provide an end for action at all. You may, for example, be proud of your achievement, sentimentally long for your former lover, or grieve over your mother's death, while at the same time lacking an end for action" (Döring 2003, pp. 219–220). While Döring grants that emotions do have representational contents, those contents are not always the cause of emotional action. Instead, she suggests the following: "It is the emotion's affect which gives it motivational force, rather than any desire being 'part' of it. Unlike a desire, an emotion's affect can still move its subject to act even if it is not necessary or actually impossible to change the world in such a way that it fits the emotion." (p. 224) Thus, she claims that the emotion's affect is unlike a desire, since it does not dispose the agent to change the world in specific ways. Nevertheless, affect can motivate actions on her view.

Scarantino and Nielsen (2015, pp. 2982–2983) point out a problem for this view: it does not explain why the felt, affective dimension of emotions is essentially motivating. For example, "the feeling of a ceiling fan blowing air through one's hair" (2015, p. 2982) does not motivate. We are left wondering then why emotional feelings/affect are essentially motivating when other feelings and affective experiences appear not to be. Another lacuna in Döring's account concerns the difference between the motivational force of feeling/affect and the motivational force of goals. How do these differ, if at all? If for example, an emotional affect is pleasant to an agent, why not think that the agent's goal is to increase or maintain the pleasant experience? In that case, the motivational force of the affect reduces to a goal of the agent. If questions like these are left unanswered, Döring's explanation of emotional actions is incomplete and unsatisfying.

Similarly, other accounts of emotional action rest on an intuitive but unexplained distinction between goals and other motivational states. Kovach and De Lancey (2005) approach the problem of emotional action from the perspective of empirically driven theories of emotions, such as basic emotion theory (e.g., Ekman 1999). They suggest that many emotional actions are explained by what they call M-emotions:

> The concept of an M-emotion is not the concept of a desire to achieve a goal or outcome state. It is rather the concept of a mental state involving a specialized motivational bodily state that has the function of producing certain behavioral responses. Thus, as an M-emotion, fear does not motivate an agent by rep-

> resenting the goal of avoiding danger. Rather, fear motivates by producing a behavioral response, such as flight." (p. 115)

To me, it seems plausible that emotions are states that have "…the function of producing certain behavioral responses" or that "…fear motivates by producing a behavioral response, such as flight." Nevertheless, I could say exactly the same about the *goal* of avoiding danger. Of course, Kovach and De Lancey clearly mean to say that fear produces "a behavioral response" in a way that is distinct from a goal representation. Perhaps they mean to say that the emotion's behavioral response does not have a goal for the agent and instead has only an evolved function. Nevertheless, they do not indicate the difference between an agent's goals and the evolved functions that otherwise direct their behavior. We are left wondering exactly how an emotion produces behavior differently from an agential goal and exactly how a "motivational bodily state" differs from a goal of the agent.

3 The Nature of Goals

This suggests the following: explanations of emotional action that do not appeal to goals must draw on a clearer distinction between goals and other kinds of motivational states. Importantly, Smith, a prominent defender of goal-based explanations of emotional actions, gives his own rough characterization of the goal concept (1987, 1998). By clarifying and refining this account, I will outline the internal structure of systems with goals. This will provide a clear basis for determining whether emotions give rise to goals for any given account of their internal structure. Since my concern is with borderline cases (i.e., emotional actions), the relevant question is which kinds of system and behavior this account rules out. Moreover, systems theory provides a helpful way of thinking about which systems satisfy the account.

A standard tool of systems theory is to model the behavior of systems with diagrams that depict causal and informational links between different parts of a system. For instance, Fig. 1a is a model of a thermostat. Each directed arrow within the models is a signal that represents the causal influence of one process upon another. Sometimes causal influence is determined by the transmission of information. Some signals come from outside the boundaries of the system being modelled. For instance, the target temperature on the thermostat is a fixed point from outside the system: it does not vary in relation to the operations of the controller or the heating plant. As such, it is depicted as an arrow that has an origin external to the feedback loop being modelled here. This signal is usually referred to as a reference signal, or set point. Systems or processes in a model are generally represented with circles and squares. Circles are used if the transformation is simple, such as summation or multiplication. Here, the circle represents a subtraction of the temperature at the sensor from the reference temperature. The controller is set up to be activated if the difference is positive (if the sensed temperature is lower than the reference). The controller's decision about how to act based on this information may be a great deal more complex. For instance, some thermostats allow the temperature to dip a few degrees below the set point before activating or they may initiate a process that involves

starting up the furnace prior to turning on the fan. This is why the controller and heating plant are depicted as boxes rather than circles. Nevertheless, the output of even a complex process may be very simple: in the case of the heating plant, it is hot air that causes a change in the sensor's temperature reading.

Returning now to Smith's remarks on the goal concept, here is his first pass characterization: "…having φ-ing as a goal is also a state that aims to have the world fit it. It…must therefore be a disposition to realize φ-ing." (Smith 1988, p. 589) What exactly does it mean to have "a state that aims to have the world fit it"?[5] At the very least, this description rules out *open-loop systems* such as the simple clothing dryer in Fig. 1b: once started, it will run for the set time, regardless of whether the clothes are dry. In other words, it is not sensitive to feedback concerning the dampness of the clothes. If we suppose that the clothing dryer's "goal" is to dry the clothes, we can ask whether it has a state that aims to have the world fit it, e.g., a state that represents dry clothes. The clothing dryer clearly does not. Its operation has the effect of drying the clothes, but it has no internal state that represents this outcome or reliably leads to its realization (without overshooting or undershooting). Another way of putting this point is that we can easily explain the dryer's behavior without positing a goal.

By contrast, thermostats and other simple feedback mechanisms are closed-loop, in the sense that they are designed to shut off or modify their behavior in response to feedback from the environment. These simple feedback systems are designed to

[5] So far, Smith's description is mostly metaphorical. For a recent criticism of the direction of fit metaphor, see Frost (2014). On Frost's view, Smith is not ultimately committed to the metaphor, because, as I discuss below, he replaces the metaphor with a functional characterization. In what follows, I sidestep recent discussions of the direction of fit metaphor. There are several reasons for this: First, most of the problems identified for the metaphor concern its usefulness for distinguishing beliefs from desires and other pro-attitudes (Milliken 2008; Schueler 1991; Sobel and Copp 2001). My interest is in whether emotional actions are goal-directed, but an argument to this conclusion need not be threatened by a permissive account of goals that fails to rule out some beliefs. If I show that emotions do not have goals on such an account, then I will have proven more, not less, than I set out to prove. So it matters little whether states of other kinds are adequately distinguished by having the opposite direction of fit from goals. Second, most discussions of direction of fit so far are criticisms of Smith's broader account of action (not his account of goals). Whether these criticisms are correct seems irrelevant to my purpose here. If my arguments concerning emotional action are correct, then they can serve as a basis for criticizing Smith's account of action (as I do in Sect. 5) by showing how some behaviors constitute actions even though they plausibly have no goals, even on his own characterization of goals. Third and most importantly, none of the counterexamples to Smith's dispositional account of belief and desire undermine its applicability to the goal concept (Milliken 2008; Schueler 1991; Sobel and Copp 2001). Consider for example, Sobel and Copp's (2001) fair-weather fan example: Sue wants the 49ers to do well, but if they begin to perform poorly, she starts wanting another team to do well. In this case, it certainly seems right that Sue has a desire that does not tend to endure when she perceives the negation of its content. However, it does not seem right to say that Sue has a goal with this feature. At the very least, it seems Smith could easily address this worry as applied to goals by pointing that a goal tends to persist in the perception of not-p unless overridden by another goal of the agent (e.g., the goal of rooting for a successful team). But in that case, the susceptibility to be overridden is not a tendency of the goal under consideration (e.g., the goal of helping her team to do well) but rather a tendency of another goal (and indeed, part of its tendency to make it the case that p). Thus, Sobel and Copp's case seems to me a problem for reducing desires to goals, but not a problem for characterizing goals in terms of the conditions and constraints discussed in this section.

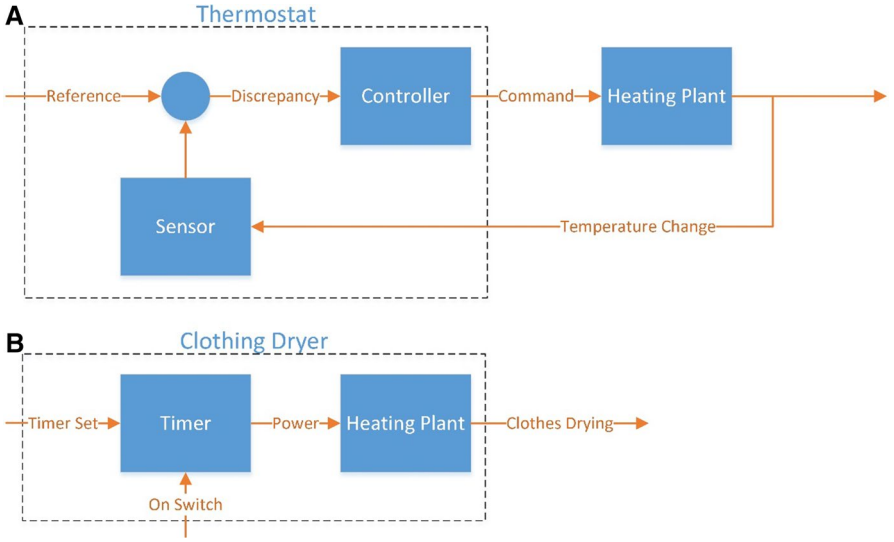

Fig. 1 **a** A systems theoretic model of a thermostat. The thermostat is a closed-loop, because it responds to feedback. **b** A systems theoretic model of a clothing dryer. The clothing dryer is open-loop because it does not respond to feedback

bring about a certain kind of outcome and to approach that outcome in a minimally flexible way. For closed-loop systems, it is difficult to see why they would not satisfy the description: typically, a thermostat has a state (its set point) that aims to have the world fit it (by getting the actual temperature to match it); this state is also a disposition to realize temperature maintenance.

If Smith had said nothing more about goals, we might characterize his view as a *deflationary* account of goal-representations. On this kind of account, goals are constituted by any state of a system that carries information about whether the system's behavior has served its function.[6] Such deflationary views are obviously insufficient for a theory of action. If thermostats count as having goals, then the class of goal-directed behaviors will encompass many things that we do not count as actions, such as pupillary movements, involuntary thermoregulatory mechanisms (e.g., for sweating and shivering), perhaps even the autumnal defoliation of deciduous trees. But Smith's account of action is clearly intended to distinguish actions from mere behaviors and passive movements of these kinds. It clearly is not meant to explain pupillary dilations/contractions in terms of motivating reasons for action, even though these bodily movements clearly have a function. Moreover, it seems implausible that such a function can partially constitute an agent's reason for acting. We can say that my pupils contract and dilate to maintain optimal light exposure to the retina (among other things), but we are never

[6] Cf. Orlandi (2014, p. 12).

in a position to say that maintaining optimal light exposure is *my own reason* for moving my pupils. After all, it is not I that moves them (Frankfurt 1978, p. 159).

Perhaps Smith can say that the problem with these behaviors is that they are properly attributed to parts of the agent rather than the agent herself. Perhaps if we focus on the parts by themselves, we will see that they do in fact have goals. This response does not forestall the suggestion that thermostats (etc.) have goals, which seems incredibly implausible. The thermostat's marker of the "desired" temperature is surely a derived representation: it is not a representation of a desired outcome for the system itself but rather for its user or designer. It simply is not a representation of the thermostat's own goals, especially qua reasons for action. We can see this by pointing out that there is no added explanatory value in calling the set point a goal or in referring to a thermostat's "reasons for acting." While the thermostat has a "reason" or "goal" from the design perspective, no explanatory aim is advanced by dropping the scare quotes. In a rigorous descriptions of how the thermostat works, talk of goals is mere metaphor.

Accordingly, there is an emerging consensus that for organisms or organismal parts to have representations in a more robust sense, posited representations must have explanatorily robust accuracy conditions or perhaps satisfaction conditions in the case of conative representations (Burge 2010; Morgan 2018; Morgan and Piccinini 2018; Orlandi 2014, pp. 7–11; Ramsey 1997). In other words, the accuracy conditions for a representation must play an indispensable role in explaining the behavior of a certain type of system. Moreover, the indispensability of accuracy conditions depends on the explanatory project of interest. The accuracy conditions of circadian clocks may play an indispensable role in explaining how plants behave (cf. Morgan 2018), but goal representations in Smith's sense are supposed to be indispensable for explaining a completely distinct phenomenon: action as distinct from mere happenings or mere behaviors (e.g., spinally mediated reflexes). Accordingly, I assume throughout this discussion that question of explanatory robustness or indispensability turn on the explanatory aims of a philosophical theory of action of the sort that Smith (1987, 1998) defends.

Consequently, any reasonable interpretation of Smith's account requires a more *committal* notion of goal representation. In other words, postulating such a state must play a deeper explanatory role than merely carrying information about whether the system's behavior has served its function. So what other explanatory work must a goal representation do to earn its keep? Though Smith does not explicitly address this question, he does say more about the functional role of beliefs and desires. Since Smith argues that goals are conceptually linked to desires, we can apply to goals much of what he says about desires:

> …the difference between beliefs and desires in terms of direction of fit comes down to a difference between the counterfactual dependence of a belief and a desire that *p*, on a perception that *not p*: roughly, a belief that *p* is a state that tends to go out of existence in the presence of a perception that *not p*, whereas a desire that *p* is a state that tends to endure, disposing the subject in that state to bring it about that *p*. Thus, we may say, attributions of beliefs and desires require that different *kinds* of counterfactuals are

> true of the subject to whom they are attributed. (Smith 1987, p. 54 emphasis in original)

Does this dispositional account rule out the thermostat having desires and thus goals? It does not. The thermostat clearly does satisfy this description, so it likewise fails to identify the more robust explanatory job that goals are supposed to fill. Like a desire, the mismatch between the thermostat's target temperature and its reading of the room's temperature tends to endure, disposing the thermostat to bring about the target room temperature. Moreover, like a belief, the thermostat's temperature readings tend "…to go [out] of existence in the presence of a perception that not p…" (Smith 1987, p. 54)[7]

Yet Smith understands this account in a more explanatorily robust manner. He thinks actions are explained by motivating reasons, and motivating reason are constituted by beliefs and desires *together*. Moreover, the relevant beliefs need to connect up with desires in content specific ways: "We understand what it is for someone to have a motivating reason at a time by thinking of him as, inter alia, having a goal at that time; the 'alia' here *includes having a conception of the means to attain that goal*." (Smith 1987, p. 54 emphasis mine) To have a conception of the means to attain a goal requires that one's conception match the goal's content. Moreover, it is because the content of these states (goal and conception of means) match that they co-constitute a reason for action on Smith's view.[8] For our purposes, this suggests that Smith is committed to a necessary condition on goal-directed action[9]:

Behavior selection constraint: if a sequence of behavior is explained by a putative goal of φ-ing, then the agent's selection/execution of the behavior depends on the agent having information that the behavior will bring the world closer to a state that realizes φ-ing.[10]

This follows because the agent's reason for acting is also why the agent selects the behavior she does. Moreover, if her reason for acting is to depend (constitutively) on a conception of the means to attain a goal, then it also minimally depends on information that a given behavior is a means to the goal. Finally, if a goal is a state that aims to have the world fit it, then this is plausibly because it constitutes an ongoing disposition to bring the world closer to a match with the goal state.

If we reconsider the thermostat, it is hard to find any reason to suppose it has this kind of information. Does the thermostat really have information that turning on the

[7] Perhaps Smith can just point out (and reasonably so, given his interests) that thermostats do not perceive anything. In that case, to specify which systems have goals, we would need to spell out which systems have perception. This would be an interesting way of proceeding but unfortunately not one that I can explore here. Thanks to Alexander Morgan for pointing out this avenue of inquiry.

[8] Cf. Bermúdez (2005, pp. 75–81)

[9] Notice the shift from a necessary condition on having a goal to a necessary condition on goal-directed behavior. The former is more difficult to specify in connection with behavior, given that a system can have goals that it never acts on. Here, I am concerned with the simpler task of saying whether a behavior sequence is produced by a (posited) goal as opposed to some other state.

[10] It is tempting to think that such information is represented as means-ends beliefs, but yielding to such a temptation threatens over-intellectualizing action. Humans and other animals likely represent means-ends information in other forms, such as non-conceptual representations of affordances.

heater will get the room closer to the correct temperature? No, it is rather that the structure of the thermostat has built in this assumption, and it cannot represent otherwise. Thus, if the forgoing is correct, it shows that the thermostat's behaviors are not governed by the goal of heating the room.

This derives from a critical feature of goals according to Smith: they serve as criteria for selecting from a range of behaviors. When a behavior is explained by a goal, this is because selection of that behavior depends on the agent's recognition that the behavior will bring the world closer to matching the goal representation. But this is just to say that the goal serves as a kind of standard for determining which behaviors are selected in its service. Herein lies a great deal of the explanatory power of goals. They explain not only behavior that is selected in one particular instance, but also in a host of counterfactual instances.[11] If I have a goal of drinking cold beer and know that there is beer in the refrigerator, the goal explains my walking over to get the beer. However, if I were to have the same goal under different conditions (e.g., I know that there is no beer in the refrigerator), it would explain a completely different sequence of behavior. In either case, the goal provides the *behavior-selection criteria*, or equivalently, the standard for determining which behaviors to select in its service.

Importantly, goals also provide the *satisfaction condition* for behavior or equivalently, the condition in which the behavior realizes the goal. We can see this by drawing out a critical implication of Smith's account. If a behavior is truly governed by a goal and if goals influence behavior by disposing an agent to bring about a state of affairs that matches the goal, then our goal attributions should be constrained in the following way: once the agent knows that the state of affairs has come about, the behavior that follows is unlikely to be explained by the relevant disposition. In other words, goals ordinarily constrain behavior in the following way:

Satisfaction constraint: if a sequence of behavior is explained by a putative goal of φ-ing, then *ceteris paribus* the behavior should cease once the agent registers that φ-ing has been realized.[12]

This constraint holds because goals determine the satisfaction condition for behavior. Once the goal is fully realized, subsequent behavior is not selected in service of that aim.[13] This feature of goals also holds a great deal of explanatory power. It explains not only why behaviors actually cease or persist but also why they *would* cease or persist in a range of counterfactual conditions.

In sum, when a goal representation guides behavior, it does so by determining both the selection criteria and satisfaction conditions for that behavior. In other

[11] This is of central importance for a theory of goals and of goal-directed behavior. For instance, Woodfield (1976, pp. 92–102) criticizes behaviorist theories of goal-directed action on the basis that they cannot explain this critical feature of goals.

[12] By contrast with Russell's famous account of goal-directed behavior (Russell 1922, Chapter 3), this constraint is predictive rather than being conceptually necessary. The ceteris paribus is meant to allow for certain exceptional cases, in which an agent has information from which she could infer that the goal has been achieved, but does not draw the relevant inferences. To me, it seems safe to say that the cases of emotional action I discuss above are not exceptional in this way.

[13] Note that some goals are never fully realized, such as the goal of maintaining optimal physical fitness.

words, the same criterion helps determine which behaviors will be selected and when the overarching pattern of behavior (in service of the goal) will cease or persist. While these distinct job descriptions are implicit in Smith's discussion of goals, he does not distinguish them. Moreover, once you mark the difference between the two, it is possible to imagine systems that have distinct criteria for selecting behavior on the one hand and determining when behavior will cease or persist on the other.

This in turn opens the possibility that other pro-attitudes can motivate action besides goals/desires. Smith himself admits that desires might not be the only states that have the right direction of fit for motivating action: "if desire is not a suitably broad category of mental state to encompass all of those states with the appropriate direction of fit, then the Humean may simply define the term 'pro-attitude' to mean 'psychological state with which the world must fit', and then claim that motivating reasons are constituted, *inter alia*, by pro-attitudes." (Smith 1987, p. 55)[14] Here, Smith assumes that other pro-attitudes can then be explicated in terms of goals. But that assumption is only valid on the supposition that the direction of fit metaphor must always be realized by a unified state "with which the world must fit." However, if motivational systems can have distinct behavior selection criteria and satisfaction conditions, then there may be no single state that uniquely identifies the changes that must take place for world to "match" the system's aim. This raises the question: what would it look like if some human behavior were controlled by such a system?

4 Simple Feedback Mechanisms and the Structure of Some Emotions

I would suggest that emotional actions (in addition to appetitive actions) are a promising answer to this question. To make good on this suggestion requires an account of emotions that distinguishes them from goals. My aim in this section is to develop such an account inspired by models of simple feedback mechanisms like thermostats. For example, we can imagine that fear is a system monitors the nearness of various threats and includes a positive feedback loop that functions to increase the distance from those threats. Similarly, we can imagine that anger monitors (among other things) the discrepancy between, on the one hand, the goodwill others manifest toward oneself and the level of goodwill that one *expects* others to manifest (Sell 2011; cf. Strawson 1963) and includes a negative feedback loop that functions to diminish the discrepancy through confrontation.[15] I develop this account below along evolutionary lines: some emotions are systems for behavior control that,

[14] There is reason to doubt whether Davidson would define pro-attitudes as "psychological states with which the world must fit." Unlike Smith, Davidson does not explicitly aim to reduce reasons for action to goals. So it seems open to him to deny that pro-attitudes involve goals. In that case, the view of emotional actions I present below may be in line with a broadly Davidsonian (or even a Humean view).

[15] These are undoubtedly oversimplifications of anger and fear, which even in nonhuman animals may actually incorporate several feedback systems. Caroline and Robert Blanchard, for instance, theorize that anti-predator responses in rats are controlled by competing aims of threat detection and avoidance, although they do not describe these competing aims in terms of feedback systems (see e.g., Blanchard and Blanchard 1987)

because of their evolutionary history, often have distinct criteria for selecting behavior on the one hand and on the other, determining when their aim has been satisfied.

As such, I begin with an influential evolutionary theory of emotions—basic emotion theory (Ekman 1999; Ekman and Cordaro 2011; Izard 2007; Tracy and Randles 2011)—which will help lay the groundwork for my own feedback model. According to basic emotion theories, each basic emotion is an evolved response to a distinct basic life problem that was recurrent in our ancestral past: among others, avoiding bodily harm (via predation or falling from great heights), avoiding poisons and parasites, and dealing with resource competition and various other social interaction problems. The solutions to some of these basic life problems can be understood as various ways of adjusting or maintaining the organism's relation to its environment: keeping distance from dangers, avoiding contact with disease vectors, maintaining some control over resources and conspecifics, etc. As such, one can say that each emotion was selected to solve a basic life problem by implementing a *relational aim*. In other words, the relational aim of an emotion is its function, and as such it is akin to the designer's goal for a thermostat. Just as we can say of the thermostat that it aims to keep the temperature at the set point because that is its function, we can say of basic emotions that they aim to keep the organism within a range of relations to its environment because that is their function. Moreover, just as we will not say that the thermostat's *own goal* is to keep the temperature at the set point (see the arguments in Sect. 3), we should not say that the goal of our emotions (much less ourselves) is to keep us within a range of relations to our environment. As I will suggest, the satisfaction conditions of emotional behavior approximate the relational aims of emotions, but these aims may not ever become goals of a given agent.

Most proponents of basic emotion theory make three important conjectures about basic emotions. First, many basic life problems concern sociality, and so many emotions solve these problems with a signaling mechanism (see e.g., Shariff and Tracy 2011).[16] Second, basic emotions are supposed to solve basic life problems by preparing the organism for certain types of action (e.g., fight or flight), and they do so by coordinating necessary changes in physiology (Ekman et al. 1983; Levenson 1992), among other things. Third, basic emotions are constituted by at least two separate subsystems. On the output side are mechanisms for emotion production, which explain how facial expressions and physiological responses (among other things) are coordinated. These production mechanisms are usually called *affect programs*.[17] Importantly, basic emotion theorists think affect programs produce behaviors that are *ballistic* in nature.[18] This means that once triggered, an emotional response (e.g.,

[16] This is the part of the theory that originally led scientists to focus on involuntary facial expressions (which are hypothesized to function as signals) as a main line of evidence for the existence of basic emotions (e.g., Ekman 1972; Matsumoto and Willingham 2009). Nevertheless, it is obviously not true that the only function of emotions concerns sociality. For instance, disgust and fear, both considered basic emotions, have obvious individualistic functions of protecting the organism from various threats to the body's survival and proper function.

[17] See Griffiths (1997) for an early and influential philosophical perspective on basic emotion theories, which he calls "affect program" theories of emotion

[18] I owe this term to John Doris (2009, p. 71).

its facial expression and physiological response) carries out to completion with little possibility of interference from other cognitive processes, somewhat like fixed action patterns from the ethological tradition (see e.g., Thorpe 1951).[19] The putative short-term, ballistic nature of these responses makes them adequately describable by an open rather than closed loop architecture (as in Fig. 2).

On the input side of the open loop, basic emotion theory postulates *automatic appraisal mechanisms*, which have the function of quickly and efficiently detecting situations relevant to each basic life problem. These mechanisms are also thought to elicit emotions outside of the organism's conscious awareness and independently of high-level perceptual processing (of the sort that occurs in various cortical regions) (LeDoux 1998, 2012). The function of each automatic appraisal mechanism is to detect discrepancies with the relational aim for each basic emotion. For example, if the relational aim is stay away from danger, the automatic appraisal mechanism would monitor closeness to danger; if the relational aim is to maintain control over resources, the mechanism may monitor challenges from conspecifics.

Since appraisal can operate on low-level inputs (independently of high-level perceptual processing, etc.), it may only respond to *imperfect* indicators of the emotion's relational aim. For example, fear of heights can be triggered by rudimentary visual cues, which do not always correspond to any real danger of falling, as when one experiences fear while walking on the skywalk in the Grand Canyon.[20] Put in terms of systems theory, the reference signals for basic emotions are not the same as their relational aims. Instead, their reference signals are likely to be relations with the environment that reliably indicate relational aims (over evolutionarily significant time spans). For each basic emotion, the organism's evolutionary history sets the reference signal (e.g., what sensory signals tend to trigger the emotion by default), though the reference signal may be adjusted over the course of an organism's lifetime, for instance, by processes of prepared learning.[21] Finally, due to the low levels at which appraisal mechanisms can operate, emoters may have no access to the

[19] Ekman and others allow that people can learn culture-specific *display rules* that are somewhat like habits for controlling the emotional response (Friesen 1973). Nevertheless, Ekman and others maintain that it is not easy to suppress or control the emotional response *voluntarily*. It is the habitual nature of display rules that supposedly allows them to interfere with the emotional response.

[20] Cf. Gendler (2008, p. 634)

[21] This is to say that some basic emotions are likely influenced by Pavlovian learning mechanisms (see e.g. Dayan and Berridge 2014) that use an organism's individual experience to expand or contract the range of elicitors for a given basic emotion. In my manner of speaking, the reference signal for fear shifts to avoiding a larger class of entities and situations as the range of elicitors for fear increases due to Pavlovian learning processes. See Kelly (2011, Chapter 2) for an account of how simple appraisal mechanisms for disgust might have evolved to flexibly adapt to an organism's environment and thus how they explain the cross cultural variation of disgust elicitors and also their intra-cultural stability. Because of this flexibility, Ekman (Ekman 2003, p. 66) thinks basic emotions are "open programs" in Mayr's (1974) sense. By contrast with closed programs, open programs can be modified over the course of development (e.g. by learning). Moreover, they can be open in this way and also ballistic (in the manner just described). The fact that they are open programs only means that the ballistic response (or its triggers) can be modified over the organism's lifespan, not necessarily within a given emotional episode

reference signals of their stock of basic emotions.[22] This would explain why people do not always know what will set off or appease their emotional responses.

It is unlikely that many of the instances of emotional action discussed above (in Sect. 2) are explained by basic emotions. Few of those actions are ballistic in nature and many are temporally extended and obviously controlled by the agent in a way that involuntary facial expressions are not. So one point of bringing up this conception of emotion is to set aside a class of emotional behaviors that are unlikely products of anyone's agency.[23] Another point of talking about basic emotions is to introduce the most prominent and extensionally modest evolutionary theory of emotion on offer today.

However, the main point is to draw out a major shortcoming of this theory (as it is usually understood) and show how it can be overcome with a relatively minor revision of the theory: many basic life problems of the sort identified by basic emotion theory clearly cannot be solved by open-loop systems. For instance, predator avoidance obviously requires that an organism flee from a predator over an extended period of time and in a way that dynamically adjusts to the predator's movements. This kind of behavior pattern requires continuous updating with respect to the reference signal for fear, which may for instance, monitor one's egocentric spatial relation to an identifiable object in the environment. Closed-loop, or feedback systems, are thus required to solve this kind of basic life problem. To appreciate the same thought from a different vantage point, fixed action patterns are insufficient to execute the necessary response. That is, no fixed action pattern can flexibly adapt to the indefinite range of flight scenarios for which an organism must be prepared. In fact,

[22] Relational aims and reference signals as I call them are different from what have been called relational goals (Scarantino 2015; Scarantino and Nielsen 2015). First, relational goals are understood as a motivational aspect of emotions: they represent goals to be executed (or not) by rational control processes. Thus they are implemented by processes for emotion production as opposed to appraisal. By contrast, on my articulation of the basic emotion picture, reference signals are better understood as informational in nature: they are the part of the emotion that carries (imperfect) information about the organism's relation to the environment and perhaps also about the ecological success condition of the emotion's relational aim. Accordingly, they are implemented by processes for appraisal or elicitation rather than emotion production.

A second and related point is that relational goals appear to be attributable to agents. This appearance comes out when Scarantino and Nielsen speak of the deliberative phase of rational control processes in emotional actions: "In the deliberative phase, the emoter must determine whether the relational goal of the emotion should be pursued and, if so, how it should be pursued, translating the abstract goal of the emotion (e.g. attacking an opponent) into a set of situated sub-goals that achieve the abstract goal in a concrete context (e.g. picking up a bottle from a nearby table and hitting the opponent on the head with it)" (p. 2989). On this account, insofar as an emotion influences action, it does so because the agent takes on the relational goal specified by the emotion. By contrast, the relational aim of fear is not usually attributable to the agent in this way. The emoter may have no clear grasp or understanding of what elicits an emotional episode (e.g., one that incorporates concepts of "danger" or "threat") in general or in a specific instance. For example, the episode could be triggered by low level sensory stimuli or simple associations thereof. More specifically, it could be triggered by an association of a particular sound with pain (or nausea or pleasure, etc.); an association that exists because of a previous pairing of that sound with a pain (cf. Seymour et al. 2007). In cases like this, the emoter may be unaware of both the immediate cause of their fear or of the relational aim that fear serves (e.g., to help predict and avoid bodily damage).

[23] In connection with this point, cf. Goldie (2000, p. 34).

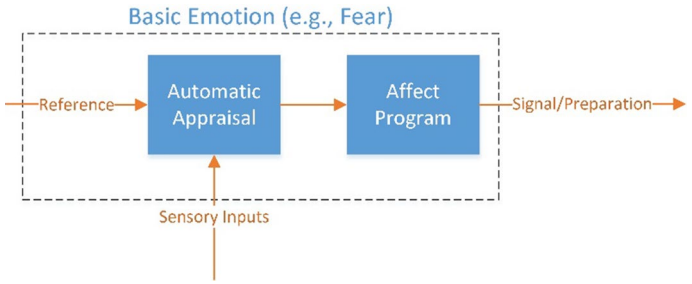

Fig. 2 Open loop structure of basic emotions

any finite, open-loop behavioral output from the emotion by itself is likely to be predictable and thus advantageous to predators rather than prey. For this and other reasons, this kind of system would be insufficient to organize a wide enough range of escape trajectories. Instead, an anti-predator system must call on a broader range of the organism's behavioral repertoire, which is plausibly controlled by other systems, ones not specialized for flight scenarios only. Moreover, to dynamically interact with these other systems in a way that guides ongoing flight behavior, the emotion must receive feedback from the environment.[24]

As such, two changes need to be made to the putative structure of basic emotions to reflect this kind of closed-loop mechanism. First, in order to more flexibly guide behavior, some evolved emotions need to interact with behavior selection systems that are more domain general and can organize a more general array of bodily movements: what I will call *executive/motor control* processes (as in Fig. 3).[25] Second, the kind of output from the affect program that seems required for this flexibility cannot be a specific motor command—for instance, of the sort that would activate certain facial expressions. Rather, the output must be more like *action tendencies*.[26] Nico Frijda introduces actions tendencies in the following way.

> Action tendencies are states of readiness to execute a given kind of action… defined by…[the] end result aimed at. One action tendency is readiness for attacking, spitting, insulting, turning one's back, or slandering, whichever of these appears possible or appropriate at a given moment… Action tendencies are hypothesized…for theoretical reasons: to account for latent readiness and to account for behavioral flexibility.''(Frijda 1986, p. 70)

[24] Scarantino (2014, 2017) makes a similar point concerning the flexibility of basic emotions

[25] I will not assume here that executive/motor control processes are identical to an agent's rational control over their actions. Some motor control processes are almost certainly beyond the agent's rational control or awareness. Cf. Dretske (2006) and Pacherie (2008)

[26] As a result, my use of the theoretical term, "affect program" is not continuous with its use in basic emotion theory. For instance, Ekman does not think that affect programs include innate action tendencies of the sort I posit here (see Ekman 2003, p. 268 n. 8). Nevertheless, I find this a useful way of picking out the subcomponent of emotions that is responsible for emotion production.

Accordingly, I will refer to action tendencies as commands sent to executive/motor control processes, which then decide how best to carry out the command (cf. Scarantino and Nielsen 2015). In the case Frijda discusses above, the command is to "move against" or to remove an obstacle (Frijda 1986, p. 88; Frijda et al. 1989, p. 214), and executive/motor control processes can then decide how to carry out this command (e.g., attacking, spitting, slandering). Nevertheless, I will not assume that action tendencies directly carry information about the reference signal of the emotion, much less its relational aim. Concerning the reference signal, this is because the reference signal concerns the appraisal mechanism, which may store information that the affect program cannot access.[27] Concerning the relational aim, it is an evolved function of which the organism may have no conception.

Emotions that meet this description I will call *simple emotions*.[28] The point of focusing on these emotions is that, like the feedback structures of simple thermostats, they were designed (or rather, selected) to bring about certain ends (or relational aims) of which the broader system need have no clear conception. Moreover, this is not a bug or glitch concerning simple emotions. Rather it is a design feature. Herein lies the significance of the evolutionary component of the theory. If every organism needed to have a structured conception of the outcomes necessary to bring about survival and reproduction, Earth would still be a barren landscape. We can say that the action-tendencies of simple emotions are *built-in strategies* for maintaining their relational aims to solve a given basic life problem. As such, these strategies are like the thermostat's built-in "assumption" that by toggling the heating plant it will accomplish the aim for which it was designed. A related design feature of simple emotions is that they do not need any complex machinery for generating expectations about the outcomes of their prepared strategies (as the behavior selection constraint would require). They can instead depend on invariances in their environment that make the prepared strategies generally effective for their relational aims.

Regardless, to explain a broad range of emotional actions in this way requires substantial supplementation with independently plausible suppositions about each emotion. Moreover, in keeping with a broadly evolutionary picture of simple emotions, these suppositions are plausibly biological in nature and concern how a given emotion solves certain basic life problems. As indicated above, the basic life problem for each emotion partially determines which relational aim a given emotion will

[27] When someone has a panic attack on an airplane because they are afraid of flying, it may be because the appraisal mechanism for fear is triggered and also the action tendency to escape or avoid a threat. Nevertheless, there may be no clear object from which flight/escape is directed. One way of explaining this is by supposing that appraisal mechanisms do not share information with affect programs about what triggered the emotion. This possibility may also help explain certain features of displaced aggression discussed below.

[28] These features of the model give it a strong resemblance to the Motivational Theory of Emotions (Scarantino 2015). Perhaps it is even an instance of that theory, though see fn. 19 for what may be a substantive difference between the two models. Perhaps the most important contribution of this model above and beyond the Motivational Theory of Emotions is the explicit recognition that the behavior selection criteria and satisfaction conditions of emotions can come apart. Regardless, the Motivational Theory of Emotions has likely influenced my thinking on this topic in more ways than I am able to trace in this paper.

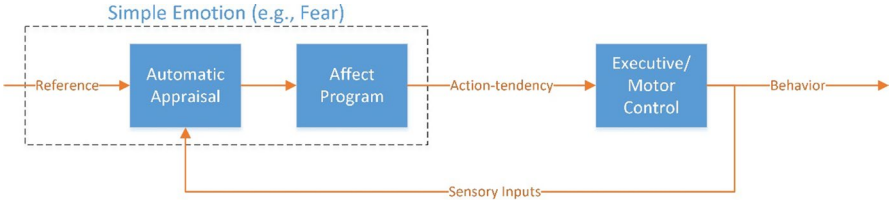

Fig. 3 The control structure of simple emotions

maintain (or change) and additionally determines which action-tendencies an emotion will rely upon to do so.[29]

We are now in position to see how simple emotions can lead to actions without goals. Consider anger and displaced aggression. Anger plausibly evolved for either of two purposes. First, it may have evolved to motivate organisms to overcome obstructions to their goals. Second, and likely later in our evolutionary history, it may have evolved (or may have been modified) to deal with resource competition among conspecifics. These two functions are related: resource competition occurs precisely when one's goal of controlling a resource is blocked by a conspecific.

In the case of humans, resource competition is extremely abstract. So, on a prominent evolutionary theory, human anger not only monitors one's control over resources and goals, but also over the disposition of others with respect to one's resources and goals (Sell 2005, 2011; Sell et al. 2009). For instance, this theory intends to explain why one might become angry if someone merely indicates through their negligence that they do not care about one's well-being, as when someone allows their dog to defecate on one's property without cleaning it up. As such, *recalibrational* theories of anger and other emotions suggest that humans monitor the degree to which others value one's welfare via a continuously updated, internally represented variable called a welfare tradeoff ratio (WTR).[30] More specifically, person A's WTR toward person B is a measure of the benefits that A would accept by imposing a cost on (or withholding a benefit from) B. If this story is right, then it may be that reference signals for anger (sometimes) concern the WTRs of other agents.[31] In that case, anger would be elicited by evidence that someone else's

[29] I say "partially" because emotions are likely to be shaped not only by the forces of natural selection in response to a basic life problem but also by phylogenetic constraints imposed by the organism's ancestry. See e.g. Griffiths (2006, 2007). This is why different organisms respond very differently to the same basic life problem. For example, given their differences in size (relative to predators), body plan and ecological niche, the moose and the mouse are likely to respond to predators in extremely different ways.

[30] Recalibrational theories make a wide range of accurate predictions concerning, among other things, punishment, revenge, forgiveness, and confession. See for example, the work of Aaron Sell, Leda Cosmides, and John Tooby and others (McCullough et al. 2012; Petersen et al. 2010, 2012; Sell 2005, 2011; Sell et al. 2009).

[31] I am inclined to think that anger has a single function of dealing with goal obstructions, and that is why it is so closely connected with violated expectations of reward or non-punishment (Berkowitz 1989, 2012). On this view, WTRs should be understood in these terms: I expect to be "rewarded" (or not punished) with a certain quality of will from others, and those expectations are violated when I register their ill-will (for discussion see, e.g., revoked reference).

WTR is lower than the reference (i.e., "desired" WTR), or equivalently, evidence that someone else would impose a cost on (or withhold a benefit from) oneself for a relatively small benefit.

Cases like Mr. Farrington's and cases of redirected aggression more broadly suggest that the appraisal mechanisms for anger do not convey to the affect program precisely *which* person's WTR is too low or *which* goal is being blocked.[32] Instead, once triggered, the affect program sends a command to executive/motor control processes to "move against" someone (or something) that may be blocking one's goals (cf. Frijda 1986, pp. 88–89; Frijda et al. 1989, p. 214). If so, then the resemblance to the thermostat appears to be in place. The anger mechanism can be elicited by one person's low WTR and thereby trigger an ongoing action tendency to move against antagonists generally. Nevertheless, the ongoing action tendency may be insensitive to whether moving against a certain person (such as Farrington's son) will satisfy the emotion by resulting in an appropriate match with the reference signal. If this feedback system adequately represents anger's influence on action, it is clear that the system does not assess the efficacy of the action tendency for achieving the relational aim of the emotion (i.e., recalibrating low WTRs or overcoming goal obstructions). Therefore, the agent's action under the influence of anger would not be guided by a goal, since he is driven to match the reference signal without any expectation that his behaviors will accomplish this end. Neither does the system have a unified standard for selecting behavior and determining when anger has been satisfied.

We can see this more clearly by considering the kind of feedback structure necessary to realize such expectations. In systems theory, this kind of information is introduced in the form of an added component, a forward model (as depicted in Fig. 3). The forward model is set up to receive a copy of commands from the controller (also known as an efference copy, depicted in Fig. 4 as the arrow from the executive/motor control process to the forward model). Together with a model of the system itself (e.g., the thermostat) and the external environment, the forward model generates the expected outcome of the command. Moreover, the expectations of the forward model are compared to actual outcomes and this comparison can be used to modify subsequent behavior. Henceforth, I will call feedback systems and emotions *complex* if they include added components like forward models).[33] Focusing back on the case of anger, a forward model could generate expectations concerning the effects of the action tendency for recalibrating the WTR of a specific antagonist, and behaviors could then be selected as instrumental for that goal. So, on this model of anger, angry actions would require something like a forward model to have a goal.

[32] This might be a feature of anger rather than a glitch. Evolutionary psychologists are keen to point out cases in which aggression is sensitive to reputational benefits (Kurzban et al. 2007). There is even some evidence in primates that displaced aggression (also known as "redirected aggression") deters subsequent aggression (Aureli et al. 1992).

[33] This is a simplification. Inverse models and some form of practical reasoning are probably also required (see, e.g., Pacherie 2008, pp. 191–194), but a forward model captures the lion's share of the explanatory burden here. A Kalman filter is another component that can play a similar role to a forward model. See e.g. Grush (2009) for further discussion.

One virtue of this model is that the function of a forward model can easily be transposed to person level psychological states. This provides us with a clear picture of how people can have, and act upon, emotional goals. The requirements are first, that a person is aware of their current emotional state (in some sense) and of the way it is impelling them to act (i.e., its behavior selection criteria).[34] This roughly corresponds to the information carried by an efference copy to a forward model. The second requirement is that a person has a sense of what a given emotional state aims at (i.e., its satisfaction condition), most likely gained through experience dealing with that emotional state. This roughly corresponds to the information that a forward model uses to generate an expected outcome of the action (i.e., the degree to which it will appease the emotion). If these requirement were met, one would know that, for example, anger is impelling her to confront someone about a specific provocation. She would also know that her anger aims at adjusting the way the person is disposed to act toward her. This person-level knowledge of one's anger can bridge the gap between the behavior-selection criteria of the emotion and its satisfaction condition, allowing an angry agent to organize her behavior around a unified goal.

This kind of awareness and knowledge does not come easily, nor is it a simple matter to act on it. If correct, this model would vindicate Aristotle's insight that appropriately tuned emotional responses require practical wisdom. On this model, practical wisdom is also required for *integration* with one's emotions. Without that wisdom, one is driven by action-tendencies of which one is not fully aware toward ends of which one has no conception. With it, one can refine or channel one's emotional impulses toward an end that they explicitly avow.

In sum, on the feedback model, emotional actions sometimes lack goals. This happens because emotions have action-tendencies that function as prepared strategies for maintaining certain relationships with their environment. By design, they do not require any conception of how the prepared strategy maintains the relevant relationship with the environment nor, short of that, a conception of what will satisfy the emotion. Nevertheless, emotional actions can have goals. Whether they do depends on the agent's ability to bridge the gap between the emotion's action tendency and its satisfaction condition.

5 Goal-Based Theories of Action

Time to take stock. I began by trying to fill a lacuna in recent attempts to explain emotional actions without reference to goals, but what of goal-based accounts of action? On this cluster of views, actions just are behaviors that are explained (appropriately, and among other things) by the agent's goal or equivalently, by the agent's desire to bring about some end. This makes goals necessary for action: if one can

[34] One way to implement this would be for the person to have a concept of what satisfies a given emotion. Such a concept would be similar to a response-dependent concept (e.g., the OUTRAGEOUS, or "that which elicits outrage"), but instead of capturing that which emotions respond to, it captures that which satisfies emotions (e.g., RETRIBUTION or "that which satisfies outrage").

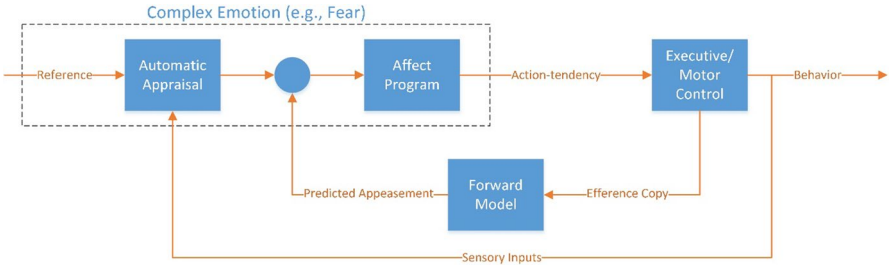

Fig. 4 A systems theoretic model of a complex emotion. The forward model makes predictions about the effectiveness of behavior for appeasing the emotion

accurately explain a behavior without reference to a goal (or perhaps, without reference to a physical state that realizes a goal), then it was not an action, on this view. For example, Smith's (1987, 1998) famous argument for the Humean theory of motivation begins with the conceptual claim that having a goal is necessary to have a motivating reasons for action. Moreover, it is clear that on Smith's view, there are no actions without motivating reasons. His own remarks about goals, discussed above, are intended to clarify and support these claims.

Thus, emotional actions present an obvious problem for this style of Humean theory; one that Smith and others have tackled explicitly: to account for Hursthouse's examples of so-called arational actions, Smith (Smith 1998, pp. 21–23) gives standard belief-desire explanations. For instance, in the case of the man rolling in his wife's clothes out of grief, he claims that "The man is doing what he is doing because he desires to roll around in his dead wife's clothes and believes that he can do so by doing just what he is doing: that is, by rolling around in those particular clothes that he is rolling around in." (p. 22). Whereas Hursthouse and others assume that the explanation of this action must invoke grief (i.e., candidate explanations she considers involve desires to express one's emotions), Smith suggests that grief is dispensable. On his view, the emotional explanation merely adds on to his basic Humean explanation by explaining *why* the man has the relevant desires (or beliefs). It is because he is in the grip of grief that he desires to roll in his wife's clothing, smell her perfume, sit in her favorite chair, etc. According to Smith, the emotion-based explanation of the man's actions presuppose the belief-desire explanation, not the reverse.[35]

Such explanations threaten to be ad hoc. To see this, recall the displaced aggression experiment (discussed in Sect. 2). In those experiments, provoked participants

[35] So, at least some action theorists wish to subsume emotional actions (or at least a considerable swath of them) under the class of rational actions, in virtue of their being more directly explicable in terms of beliefs and desires. Another way of addressing the problems raised by these emotional actions is to deny that they are really actions. Yet there appears to be wide agreement among the parties to this discussion that many of the emotional behaviors introduced by Hursthouse are in fact actions (Smith 1998, 21–23; Goldie 2000, 26–28, 34–37). To my knowledge, none of Hursthouse's interlocutors have undertaken to show that her entire range of examples are not actions.

evaluated the assistant's performance negatively only when they were "triggered" by her reading mistakes and her report that the participant scored below average on a task. Moreover, the provoked but un-triggered participants were no more likely to aggress against the assistant than the un-provoked participants ("triggered" or not). We can put the following question to the goal-based account: why would angered participants give negative evaluations only when triggered? If we follow Smith's (1998) line of thought, it must be that the actions of participants are explained by a specific desire, perhaps to adjust one's relative status toward someone, and a belief that by acting thusly (e.g., checking this box and that) one will bring about the desired adjustment. On this view, it is the desire itself, not the emotion, which explains the action. Put in terms of goals, the agent has a goal to bring about states of affairs in which the agent has a higher relative status, and the goal disposes the agent to perform actions of checking boxes on a performance evaluation until the goal state has been realized. However, Smith cannot say that anger always instates goals of this kind. If it did, then provoked participants would have given negative evaluations even when they were not triggered. Alternatively, if Smith says that anger only causes one to form this goal in one case but not the other, it will seem as if he is positing goals ad hoc; hand-fitting each goal attribution to explain each case. Obviously, this kind of hand-fitting procedure would make his belief-desire explanations unfalsifiable.[36] This worry is exacerbated by the fact that Smith does not suggest any principled way of predicting what goals or desires an angry agent will have in different contexts.

Smith can pursue two strategies to address this issue. First, Smith could insist that anger does instate goals with the same content but that it takes work to figure out what that content is. For example, we can improve predictive accuracy by adding that the goal involves adjusting one's relative status *toward people who give offense*. Nevertheless, it remains unclear why angered participants would consider the triggering behavior of the assistant to be offensive. More importantly, it seems implausible that the desire to adjust one's status is present in every case of redirected aggression. For instance, a case of kicking a door in which one has jammed one's finger is not easily recognized as a case of adjusting one's relative status,[37] even though it is easily recognized as a case of displaced aggression caused by anger. In other words, if Smith insists that anger gives rise to a desire with a single kind of content, then the goal needs to explain the entire range of angry actions (door-kicking, storming

[36] Scarantino and Nielsen (2015) also criticize Smith's account as ad hoc. They accuse Smith of postulating *sui generis* emotional desires that can be "elastically stretched to fit whatever properties emotions may be found to have…" (p. 2986). For example, the desires necessary to explain displaced aggression would need to differ from ordinary desires in terms of their strength and capacity to override reasoning. The criticism I make here is slightly different. Even if we assume that Smith can explain the relevant cases with ordinary desires, he still needs some principled basis for determining which kinds of desires an emotion will instate, where different kinds of desires are individuated by content rather than strength or impulsivity.

[37] I, for one, doubt that the agent is trying to put the door "in its place." However, some have defended a claim along these lines (see, e.g. Nussbaum 2016, pp. 18–19).

out of rooms, physically assaulting someone). Nevertheless, it is difficult to see how any clearly specified goal representation can explain all of these cases.

My point is not that Smith is dead to rights, but that he faces a difficult dilemma. He must either provide an account of the goals anger instates across different contexts, or he must insist that anger always instates goals with a specific content. Both horns of this dilemma becomes especially pressing in the presence of alternate explanation of emotional actions, like the feedback model of emotional action, which appeals directly to emotions without referring to goals. On the first horn of the dilemma, Smith's account must be supplemented with a principled way of constraining the desires or goals that explain emotional actions before it can even be tested against the feedback model. In other words, he needs a set of independently plausible assumptions about different emotions and the goals that they instate. Nevertheless, to specify the goals or desires toward which emotions dispose us, Smith cannot draw on the same evolutionary resources discussed above. While he *could* draw on this evolutionary picture to specify the sorts of *outcomes* to which a given emotion tends, he cannot assume that agents represent these outcomes *as goals* if this evolutionary story is right. A chief evolutionary advantage of emotions *qua* simple feedback systems is that they do not require substantial representational structure of the sort that is necessary to generate expectations of a behavior's efficacy (much less full-fledged goal representations). Therefore, to incorporate the same evolutionary considerations, Smith needs independent reasons to suppose that emotional aims are represented as goals.

On the second horn of the dilemma, the feedback model of emotions has a promising explanation for why any single goal representation fails to make the entire range of angry actions intelligible. On this model, there is no single picture of the world that anger disposes one to bring about. In every case, angry actions are guided by action tendencies that involve "moving against" some salient goal-blocker. Nevertheless, these action tendencies do not capture the overarching aim of angry behaviors, or the conditions in which anger is satisfied. If the evolutionary considerations above are correct, the aim is either to "recalibrate" the dispositions of others toward oneself or to remove an obstacle to one's goals (or both). Even in the case where the action-tendency and the relational signal appear to coincide on removing obstacles, the result is not a goal representation: nothing insures that the elicitor of one's anger is identical to the target of its action tendency, nor that the target of the action tendency is the same person (or obstruction) whose apology (or removal) would satisfy one's anger.

5.1 Objection: Action-Tendencies are Goals

In the end, it seems Smith should admit that the feedback model is one possible way in which emotional actions could be organized. Even if he does, there is a way to bring this account into alignment with his goal-based account of actions. He can object by arguing that the univocal goal of angry actions is to execute the angry action tendency: to move against a salient goal-blocker (or whichever action tendency scientists converge on for anger). A superficial benefit of this move is that

it would allow Smith to better explain the phenomenon of displaced aggression as manifested in the experiments described above. He could say that when participants are angered and triggered by a report of below average performance, they are disposed to move against any person who appears to block their goals, including their (putative) goal of performing no worse than average.[38] When participants are not triggered, the assistant has not blocked this goal. When they are not provoked, they do not have the action tendency. A second, more substantial benefit is this: it seems unlikely that executive/motor control processes could execute this action tendency (to "move against…") across varying conditions without generating expectations about what the command is intended to accomplish. For example, aggression measures in psychological experiments can take a variety of forms depending on the kind of harm participants are able to inflict on a target. These include negative performance evaluations (as in the experiment described above); the amount of hot sauce that one can force the target to consume (given that the target does not like spicy foods); and the duration and intensity of uncomfortable stimuli the target will experience (such as electrical shocks administered to the target's skin, air blasts delivered to the target's neck, or noise blasts delivered to the target's ears). Without a fairly complex set of expectations concerning the efficacy of these actions for "moving against…" the target, we would not expect to observe such varied manifestations of aggression when anger is provoked. Thus, the action tendency does appear to meet the behavior selection constraint discussed in Sect. 3. Specifically, if a sequence of behavior is explained by a putative goal of φ-ing (in this case "moving against…"), then the agent's selection/execution of the behavior depends on the agent having information that the behavior will bring the world closer to a state that realizes φ-ing ("moving against…").

However, this is only a necessary condition on a behavior being organized by a goal. So, we can ask whether the action tendency is better characterized as the goal of angry actions or as a sub-goal. In favor of the latter explanation is this: a goal of executing the action tendency is not always sufficient to explain the agent's overarching pattern of behavior while under the influence of anger, nor does it correctly specify the state of the world that would satisfy her anger and cause her angry action tendencies to cease. In this respect, executing the action tendency is more like a subgoal or intention than the goal of angry actions.

To see this, consider an example: if one's goal is to walk to a meeting across campus, this goal will be implemented by a number of situated sub-goals and motor commands, the latter of which are organized by low-level action-control subsystems. Moreover, even at the lowest level, these subsystems rely on expectations of efficacy (see, e.g., Pacherie 2008), likely generated by forward models: maintaining one's balance and gait over uneven terrain plausible requires something like this kind of control structure.[39] Yet, to say that each movement of one's legs over uneven terrain

[38] The details of the experiment suggest to me that participants would take on this goal. All participants are led to believe that a confederate got three more anagrams correct on the initial task.

[39] On the importance of forward models in motor commands, see Kandel et al (2012, pp. 744–760). Stepping movements over smooth ground are spinally mediated and therefore probably do not require a forward model. Nevertheless, walking over uneven terrain and avoiding obstacles in one's path requires

is guided by a forward model is consistent with saying that one has no overarching goal in walking down a given path (perhaps they are walking aimlessly). That is, we can distinguish between on the one hand, the low-level expectations that guide one's footsteps (or those that guide one's movement from waypoint to waypoint) and on the other hand, the higher-level expectations that monitor one's progress toward the superordinate goal of getting to the meeting. Even if one were walking aimlessly, each footstep would still be guided to a specific place on the ground and would be properly tuned to maintain one's balance and gait. Each leg of an aimless journey might still be guided toward specific landmarks.

My suggestion here is that the expectations that guide the execution of an emotional action tendency (to "move against") resemble the expectations that guide each footstep (or that move one toward each waypoint) more than they resemble the expectations that guide the walker toward her superordinate goal in walking. To say that the agent's goal is to put her foot at a specific place on the trail is an impoverished explanation at best and misleading at worst, and the same goes for the "goal" of executing an emotion's action tendency. If we postulate goals for each step along the hiker's path, we obscure the point that a different pattern of footsteps could lead the hiker to exactly the same destination. Likewise, if we postulate goals to move against this goal-blocker or that one (e.g., by checking this box in an evaluation or putting this much hot sauce on a cracker), then we obscure the point that each of these forms of aggression are plausibly aimed at recalibrating someone's WTR toward the angered agent. In other words, the agent's ongoing motivation to continue engaging in these and other aggressive behaviors (perhaps toward a number of different people or objects) may very well depend on signals that the targets' WTR toward them has been successfully adjusted, such as an effusive apology from the experimenter (see, e.g., Funk, McGeer, and Gollwitzer 2014; Gollwitzer and Denzler 2009). Consequently, the *aim* of WTR recalibration seems more appropriate than the *goal* of moving against this or that target for explaining the direction of an agent's pattern of behavior over the whole emotion episode (at least in some cases). Hence, it is a better candidate for explaining why the agent acted thusly under the emotion's influence. If so, then this overall pattern of behavior (guided by the aim of recalibration) constitutes an action that need not be explained by any goal of the agent. Again, this is because the behavior selection criteria (e.g., "moving against") of anger are encapsulated from its satisfaction condition (e.g., recalibrating a WTR).

Footnote 39 (continued)
visual guidance that appears to be implemented by the posterior parietal cortex, since lesions to this area interfere with the avoidance of obstacles (Kandel et al. 2012, p. 828). Moreover, it seems likely that the function of this area is to implement forward models of limb movements.

6 Conclusion

If these arguments are correct, then not all human action can be subsumed under the heading of goal-directed behavior. Once we flesh out the nature of goals, we can see the possibility that human action might be directed by other motivational states, such as the simple emotions I describe above. These emotions can motivate action with decoupled behavior selection criteria and satisfaction conditions and thus without any unified goal representation that captures the agent's understanding of their reason for acting. In that case, we have to reassess what makes emotional behaviors (*inter alia*) count as actions. If it is not that they are guided by the agent's goals, then what is it about them that makes them actions? I have little hope of fully defending an answer to this question here.

Instead, I conclude with some promising lines of thought that may leave us better off than we would be with a goal-based account (or at least no worse off). To see what I mean, consider again Michael Smith's (1987, 1998) defense of goal-based theories of motivation and action. Rather than attempting to explain what makes a system capable of representing the world, and its own goals in relation to the world, Smith takes for granted much of the folk understanding of psychological states—including features of beliefs, desires, and goal.[40] I see two problems with this approach. First, it holds hostage the concept of action to the fate of the psychological terms (e.g., goals, beliefs, and desires) that comprise the folk understanding of mind in action. The arguments above (in Sect. 4) suggest that folk psychology leaves out interesting and important kinds of states, such as simple emotions. If I am right that simple emotions can contribute just as well as goals to the organization of action, then clearly, goal-based accounts fail to capture the essence of action. This suggests that an account of action would fare better the less it is committed to a specific ontology of psychological states.

Second, if a goal-based account of action is to mark the difference between actions and mere behaviors, it must answer deep and intractable questions about goals (similar to those broached in Sect. 3): What makes a goal properly attributable to a system, rather than to its designer or to natural selection? What are the attributes in virtue of which a system can have goals of its own (among other psychological states)? If the goal-based account of action leaves these questions aside, then surely, it cannot give any satisfactory answers to the set of questions at which a theory of action begins: what makes a *behavior* properly attributable to a system, rather than to its designer or to natural selection (etc.)? What are the attributes in virtue of which a system can organize its *own actions*? My own view is that these clusters of questions are intimately connected: if we could discover what makes a system capable of having psychological states like beliefs, desires, goals and emotions, we could not possibly remain ignorant about its abilities to organize its own behavior.

[40] Given that his primary theoretical interests are metaethical, this is understandable.

If this is correct, then perhaps we would do better to anchor the theory of action to the nature of subjectivity rather than to the nature of goals.[41] On this kind of approach, central questions about agency concern that which makes a system the subject of psychological states, and actions are naturally understood as any bodily movements that are organized by the subject (perhaps irrespective of the subject's own grasp of the behavior's end). In other words, an understanding of subjectivity sets the standards for attributing psychological states and actions alike. This approach appears to mitigate both of the problems I see with some goal-based accounts. First, an account of subjectivity aims to capture what makes a system the subject of *any* psychological state, and thus it need not be constrained at the outset by a specific ontology of psychological states. Second, it promises a deep and unified understanding of agency, one that captures the conceptual interdependency of mind and action.

Acknowledgements In more ways than one, writing this paper has helped me understand how crucial feedback can be. It would not have reached its current form without generous help from Olivia Bailey, José Luis Bermúdez, Matt Bower, John Doris, Colin Klein, Charlie Kurth, Ron Mallon, Alexander Morgan, Brooke Robb, Elizabeth Schechter, Maura Tumulty, and numerous reviewers and conference attendees.

References

Aureli, F., Cozzolino, R., Cordischi, C., & Scucchi, S. (1992). Kin-oriented redirection among Japanese macaques: an expression of a revenge system? *Animal Behaviour, 44,* 283–291.

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin, 106*(1), 59–73.

Berkowitz, Leonard. (2012). A different view of anger: the cognitive-neoassociation conception of the relation of anger to aggression. *Aggressive Behavior, 38*(4), 322–333. https://doi.org/10.1002/ab.21432.

Bermúdez, J. L. (2000). *The paradox of self-consciousness*. Cambridge: The MIT Press.

Bermúdez, J. L. (2005). *Philosophy of Psychology: A Contemporary Introduction*. New York: Routledge.

Blanchard, R. J., & Blanchard, D. C. (1987). An ethoexperimental approach to the study of fear. *The Psychological Record*, *37*(3), 305.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Davidson, D. (2001). *Essays on actions and events*. Oxford: Oxford University Press. https://doi.org/10.1093/0199246270.001.0001.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective and Behavioral Neuroscience, 14*(2), 473–492. https://doi.org/10.3758/s13415-014-0277-8.

Dennett, D. C. (1997). True Believers: The intentional strategy and why it works. In J. Haugeland (Ed.), *Mind design II* (pp. 57–79). Cambridge, MA: MIT Press.

Döring, S. A. (2003). Explaining action by emotion. *Philosophical Quarterly, 53*(211), 214–230. https://doi.org/10.1111/1467-9213.00307.

Doris, J. M. (2009). Skepticism about persons. *Philosophical Issues, 19,* 57–91.

Dretske, F. I. (2006). Perception without awareness. In T. Gendler & J. Hawthorne (Eds.), *Perceptual Experience* (pp. 147–180). Oxford: Oxford University Press.

[41] For example, the intellectual descendants of Kant have tended to focus on subjectivity as a condition for perception, self-consciousness, and mindedness more generally (see, e.g., Bermúdez 2000; Grush 2007; Morgan 2018, pp. 5425–5426; Strawson 1959).

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation* (Vol. 19, pp. 207–283). Lincoln, NE: University of Nebraska Press.
Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *The handbook of cognition and emotion* (pp. 45–60). Sussex: Wiley.
Ekman, P. (2003). *Emotion revealed: Understanding faces and feelings*. New Haven: Phoenix Press.
Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364–370. https://doi.org/10.1177/1754073911410740.
Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*(4616), 1208–1210.
Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly, 15*(2), 157–162.
Friesen, W. V. (1973). *Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules*. San Francisco: University of California.
Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press. https://doi.org/10.1093/0199253048.001.0001.
Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology, 57*(2), 212–228. https://doi.org/10.1037//0022-3514.57.2.212.
Frost, K. (2014). On the very idea of direction of fit. *Philosophical Review, 123*(4), 429–484.
Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40*(8), 986–997. https://doi.org/10.1177/0146167214533130.
Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy, 105*(10), 634–663.
Goldie, P. (2000). Explaining expressions of emotion. *Mind, 109*(433), 25–38. https://doi.org/10.2307/2659992.
Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45*(4), 840–844. https://doi.org/10.1016/j.jesp.2009.03.001.
Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.
Griffiths, P. E. (2006). Function, homology, and character individuation. *Philosophy of Science, 73*(1), 1–25.
Griffiths, P. E. (2007). Evo-Devo Meets the Mind: Towards a developmental evolutionary psychology. In R. Sansom & R. Brandon (Eds.), *Integrating evolution and development: From theory to practice* (pp. 195–225). Cambridge: The MIT Press.
Grush, R. (2007). Skill theory v.2 0: dispositions, emulation, and spatial perception. *Synthese, 159*(3), 389–416.
Grush, R. (2009). *Space, time and objects*. In J. Bickle (Ed.), The Oxford hand book of philosophy and neuroscience (pp. 311–345). Oxford: Oxford University Press.
Hursthouse, R. (1991). Arational actions. *The Journal of Philosophy, 88*(2), 57–68.
Izard, C. E. (2007). Basic Emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science, 2*(3), 260–280. https://doi.org/10.1111/j.1745-6916.2007.00044.x.
Joyce, J. (2011). Dubliners. In S. Latham (Ed.), *Dubliners*. New York: Longman.
Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudsperth, A. J. (2012). *Principles of neural science* (5th ed.). New York: McGraw-Hill.
Kelly, D. R. (2011). *Yuck! The nature and moral significance of disgust*. Cambridge: MIT Press.
Kovach, A., & DeLancey, C. (2005). On emotions and the explanation of behavior. *Nous, 39*(1), 106–122. https://doi.org/10.1111/j.0029-4624.2005.00495.x.
Kurzban, R., Descioli, P., & Obrien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001.
LeDoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Simon and Schuster.
LeDoux, J. (2012). Rethinking the emotional brain. *Neuron, 73*(4), 653–676. https://doi.org/10.1016/j.neuron.2012.02.004.
Levenson, R. W. (1992). Autonomic nervous system differences among emotions. *Psychological Science, 3*(1), 23–27.

Marcus-Newhall, A., Pedersen, W. C., Carlson, M., & Miller, N. (2000). Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology, 78*(4), 670–689. https://doi.org/10.1037//0022-3514.78.4.67O.

Matsumoto, D., & Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology, 96*(1), 1–10. https://doi.org/10.1037/a0014037.

Mayr, E. (1974). Behavior programs and evolutionary strategies: natural selection sometimes favors a genetically "closed" behavior program, sometimes an "open" one. *American Scientist*, *62*(6), 650–659.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2012). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences, 36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160.

Milliken, J. (2008). In a fitter direction: Moving beyond the direction of fit picture of belief and desire. *Ethical Theory and Moral Practice, 11,* 563–571.

Morgan, A. (2018). Mindless accuracy: On the ubiquity of content in nature. *Synthese, 195*(12), 5403–5429. https://doi.org/10.1007/s11229-018-02011-w.

Morgan, A., & Piccinini, G. (2018). Towards a cognitive neuroscience of intentionality. *Minds and Machines, 28*(1), 119–139. https://doi.org/10.1007/s11023-017-9437-2.

Nussbaum, M. C. (2016). *Anger and forgiveness: Resentment, generosity, justice*. New York: Oxford University Press.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. New York: Oxford University Press.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*(1), 179–217. https://doi.org/10.1016/j.cognition.2007.09.003.

Pedersen, W. C., Gonzales, C., & Miller, N. (2000). The moderating effect of trivial triggering provocation on displaced aggression. *Journal of Personality and Social Psychology, 78*(5), 913–927.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In H. Hogh-Olesen (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 72–131). New York: Palgrave Macmillan.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior, 33*(6), 682–695. https://doi.org/10.1016/j.evolhumbehav.2012.05.003.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep? *Mind and Language, 12*(1), 34–66. https://doi.org/10.1111/1468-0017.00035.

Russell, B. (1922). *The Analysis of Mind*. London: George Allen & Unwin.

Scarantino, A. (2014). Basic emotions, psychological construction and the problem of variability. In L. F. Feldman & J. A. Russell (Eds.), *The Psychological Construction of Emotion* (pp. 334–376). New York: The Guilford Press.

Scarantino, A. (2015). The motivational theory of emotions. In J. D'Arms & D. Jacobson (Eds.), *Moral psychology and human agency* (pp. 156–185). Oxford: Oxford University Press.

Scarantino, A. (2017). Do emotions cause actions, and if so how? *Emotion Review*, *9*(4), 326–334. https://doi.org/10.1177/1754073916679005.

Scarantino, A., & Nielsen, M. (2015). Voodoo dolls and angry lions: How emotions explain arational actions. *Philosophical Studies, 172,* 2975–2998. https://doi.org/10.1007/s11098-015-0452-y.

Schueler, G. F. (1991). Pro-attitudes and direction of fit. *Mind, 100*(2), 277–281.

Sell, A. (2005). *Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger*. Barbara: University of California Santa.

Sell, A. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior, 16*(5), 381–389. https://doi.org/10.1016/j.avb.2011.04.013.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences of the United States of America, 106*(35), 15073–15078. https://doi.org/10.1073/pnas.0904312106.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Neuroscience, 8*(4), 300–312. https://doi.org/10.1038/nrn2119.

Shariff, A. F., & Tracy, J. L. (2011). What are emotion expressions for? *Current Directions in Psychological Science, 20*(6), 395–399. https://doi.org/10.1177/0963721411424739.

Smith, M. (1987). The Humean theory of motivation. *Mind, 96*(381), 36–61. https://doi.org/10.1093/mind/XCVI.381.36.

Smith, M. (1988). On humeans, anti-humeans, and motivation: A reply to pettit. *Mind, 97*(388), 589–595. https://doi.org/10.1093/mind/XCVII.388.589.
Smith, M. (1998). *The possibility of philosophy of action*. In J. Bransen & S. E. Cuypers (Eds.) *Human Action, Deliberation and Causation*. New York: Springer.
Sobel, D., & Copp, D. (2001). Against direction of fit accounts of belief and desire. *Analysis, 61*(1), 44–53.
Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.
Strawson, Peter F. (1963). Freedom and resentment. In G. Watson (Ed.), *Free will* (pp. 59–80). Oxford: Oxford University Press.
Thorpe, W. H. (1951). The definition of some terms used in animal behavior studies. *Bulletin of Animal Behaviour, 9,* 34–40.
Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review, 3*(4), 397–405. https://doi.org/10.1177/1754073911410747.
Woodfield, A. (1976). *Teleology*. Cambridge: Cambridge University Press.