

6 On Making Believable Emotional Agents Believable

Andrew Ortony

Abstract

How do we make an emotional agent a believable emotional agent? Part of the answer is that we have to be able to design agents whose behaviors and motivational states have some consistency. This necessitates ensuring situationally and individually appropriate internal responses (emotions), ensuring situationally and individually appropriate external responses (behaviors and behavioral inclinations), and arranging for sensible coordination between internal and external responses. Situationally appropriate responses depend on implementing a robust model of emotion elicitation and emotion-to-response relations. Individual appropriateness requires a theory of personality viewed as a generative engine that provides coherence, consistency, and thus some measure of predictability.

6.1 Making Believable Emotional Agents Believable

What does it take to make an emotional agent a *believable* emotional agent? If we take a broad view of believability—one that takes us beyond trying to induce an illusion of life through what Stern (chapter 12 of this volume) refers to as the “Eliza effect,” to the idea of generating behavior that is genuinely plausible—then we have to do more than just arrange for the coordination of, for example, language and action. Rather, and certainly in the context of *emotional* agents, the behaviors to be generated—and the motivational states that subserve them—have to have some *consistency*, for consistency across similar situations is one of the most salient aspects of human behavior. If my mother responds with terror on seeing a mouse in her bedroom today, I generally expect her to respond with terror tomorrow. Unless there is some consistency in an agent’s emotional reactions and motivational states, as well as in the observable behaviors associated with such reactions and states, much of what the agent does will not make sense. To be sure, people do not always react in the same way in the same

kind of situation—there must be variability within consistency, but equally surely there is *some* consistency—enough in fact for it to be meaningful to speak of people behaving in character. An agent whose behaviors were so arbitrary that they made no sense would probably strike us as psychotic, and Parry (e.g., Colby 1981) notwithstanding, building psychotics is not generally what we have in mind when we think about building believable emotional agents or modeling human ones.

But consistency is not sufficient for an agent to be believable. An agent's behavior also has to be *coherent*. In other words, believability entails not only that emotions, motivations, and actions fit together in a meaningful and intelligible way at the local (moment-to-moment) level, but also that they cohere at a more global level—across different kinds of situations, and over quite long time periods. For example, I know that my daughter intensely dislikes meat—it disgusts her to even think about eating something that once had a face. Knowing this, I know that she would experience disgust if she were to suddenly learn that she was eating something that contained meat (e.g., beef bouillon, not vegetable bouillon), and I would expect her disgust to influence her behavior—she would grimace, and push the plate away, and make some hideous noise. In other words, I expect her emotion-related behaviors to be consonant with (i.e., appropriate for) her emotions. But I also expect coherence with other realms of her life. Accordingly, I would be amazed if she told me that just for the fun of it, she had taken a summer job in a butcher's shop (unless perhaps I learned that she had taken the job with a view to desensitizing herself). Clearly, the issue of coherence is an important part of the solution to the problem of how to construct believable emotional agents.

6.2 Consistency and Variability in Emotions

It is an interesting fact about humans that they are often able to predict with reasonable accuracy how other individuals will respond to and behave in certain kinds of situations. These predictions are rarely perfect, partly because when we make them, we generally have imperfect information, and partly because the people whose behavior and responses we are predicting do not always respond in the same way in similar situations. Nevertheless, it is certainly possible to predict to some degree what other people

(especially those whom we know well) will do and how they will feel and respond (or be inclined to respond) under varying circumstances. We also know that certain kinds of people tend to respond in similar ways. In other words, to some extent, there is both within-individual consistency and cross-individual consistency.

So what makes it possible to predict and understand with any accuracy at all other people's feelings, inclinations, and behavior? At least part of the answer lies in the fact that their emotions and corresponding behavioral inclinations are not randomly related to the situations in which they find themselves, for if they were, we'd be unable to predict anything. But if the emotions, motivations, and behaviors of people are not randomly associated with the situations whence they arise, there must be some psychological constraints that limit the responses that are produced. And indeed, there are. Sometimes the constraints are very limiting (as with reflexes, such as the startle response) and sometimes they are less so—merely circumscribing a set of possibilities, with other factors, both personal and contextual, contributing to the response selection. But either way, there are constraints on the internal responses to situations—that is, on the internal affective states and conditions that arise in people—and on the external actions that are associated with those states and conditions.

Discussion

Sloman: You used the word "behavior" several times, and I suspect you are talking about intentions rather than behavior.

Ortony: Yes, that's why I called it motivational-behavioral component.

Sloman: But it's absolutely a crucial thing, for example, with regard to your daughter. She might well be going to work at a butcher's for the same kind of reason as somebody who belongs to a police group might join a terrorists' organization. It's the intention that is important, and the behavior might just be an appropriate means.

Ortony: Right. And actually I meant to mention this in the imaginary context of my daughter going to work at the butcher's, because one thing we would try to do to maintain our belief that people's behavior is coherent is that we would come up with an explanation, such as: "she is trying to desensitize herself." We

would not feel comfortable letting these two parts of behavior coexist—we would think that she was crazy or something.

There are two classes of theories in psychology that are relevant to these issues. Theories of emotion, and theories of personality. Consider first, emotion theories—especially cognitive ones, which are often incorporated into affective artifacts. The principal agenda of cognitive theories of emotion is the characterization of the relation between people's construals of the situations in which they find themselves and the kinds of emotions that result. The specification of such relationships is a specification of the constraints that construals of the world impose on emotional states. And these constraints are a major source of consistency, both within and across individuals. At the same time, they are only constraints—they do not come close to fully determining what a particular individual will feel or do on a particular occasion because they work in concert with several sources of variation. These are (1) individual differences in the mappings from world situations to construals (e.g., members of the winning and losing teams in a football game have different mappings from the same objective event), (2) individual differences in something that we might call emotionality (e.g., some of the team members might be more prone to respond emotionally to good or bad outcomes than others), and (3) the current state of the individual at the time (e.g., current concerns, goals, mood).

Mappings from particular types of emotions to classes of behavioral inclinations and behaviors are similarly constrained, and thus constitute another source of consistency. This is an area that only a few psychologists (e.g., Averill 1982, on anger) have studied in any very deep way, except with respect to facial expressions (e.g., Ekman 1982), although it was of considerable interest to Darwin who first wrote about it at length in his 1872 (first edition) book, *The Expression of Emotions in Man and Animals*. However, probably because the linkage between emotions and behaviors is often very flexible, there has been little effort to develop systematic accounts of it. But again, we know that the relation cannot be random, and this means that it ought to be possible to identify some principles governing constraints on the relation between what we feel and what we do, or are inclined to do. And again, whereas there are some constraining principles governing the emotion-behavior connection—principles that are the source of some con-

sistency—there are also various factors (e.g., emotionality, again) that give rise to variation.

People only get into emotional states when they *care* about something (Ortony, Clore, and Foss 1987)—when they view something as somehow good or bad. If there's no caring, there's no emoting. This suggests that the way to characterize emotions is in terms of the different ways there might be for feeling good or bad about things. Furthermore, many traits can be regarded as chronic propensities to get into corresponding emotional states. For example, an anxious person is one who experiences fear emotions more easily (and therefore more frequently) than most people, and an affectionate person is one who is likely to experience (and demonstrate) affection more readily than less affectionate people. This means that if we have a way of representing and creating internal states that correspond to emotions, we can capture many traits too. This is important because, at the level of individuals—and this is one of my main points—traits are a major source of emotional and behavioral consistency.

Many psychologists (e.g., Ortony, Clore, and Collins 1988; Roseman, Antoniou, and Jose 1996; Scherer 1997) have proposed schemes for representing the conditions under which emotions are elicited. In our own work (which in affective computing circles is often referred to as the OCC model), we proposed a scheme that we thought accommodated a wide range of emotions within the framework of twenty-two distinct emotion types. Over the years, Gerald Clore and I, together with some of our students, collected considerable empirical support for many of the basic ideas. However, for the purposes of building believable artifacts, I think we might want to consolidate some of our categories of emotions. So, instead of the rather cumbersome (and to some degree arbitrary) analysis we proposed in 1988, I think it is worth considering collapsing some of the original categories down to five distinct positive and five negative specializations of two basic types of affective reactions—positive and negative ones—as shown in table 6.1.

I think that these categories have enough generative capacity to endow any affective agent with the potential for a rich and varied emotional life. As the information processing capabilities of the agent become richer, more elaborate ways of characterizing the good and the bad become possible, so that one can imagine a system starting with only the competence to differentiate positive from negative and then developing progressively more elaborate

Table 6.1 Five specializations of generalized good and bad feelings (collapsed from Ortony, Clore, and Collins 1988)

Positive reactions

- because something good happened (joy, happiness etc.)
- about the possibility of something good happening (hope)
- because a feared bad thing didn't happen (relief)
- about a self-initiated praiseworthy act (pride, gratification)
- about an other-initiated praiseworthy act (gratitude, admiration)
- because one finds someone/thing appealing or attractive (love, like, etc.)

Negative reactions

- because something bad happened (distress, sadness, etc.)
 - about the possibility of something bad happening (fear, etc.)
 - because a hoped-for good thing didn't happen (disappointment)
 - about a self-initiated blameworthy act (remorse, self-anger, shame, etc.)
 - about an other-initiated blameworthy act (anger, reproach, etc.)
 - because one finds someone/thing unappealing or unattractive (hate, dislike, etc.)
-

The first entry in each group of six is the undifferentiated (positive or negative) reaction. The remaining five entries are specializations (the first pair goal-based, the second standards-based, and the last taste-based).

categories. A simple example of this idea is that fear can be viewed as a special case of a negative feeling about something bad happening—with the bad thing being the *prospect* of something bad happening. If one adopts this position, then one is left with the idea that the main driving force underlying all emotions is the registration of good and bad and that discrete emotions can arise to the extent that the nature of what is good and bad for the agent can be and is elaborated. Indeed, this may well be how humans develop increasingly sophisticated emotion systems as they move from infancy through childhood to adulthood.

So, specifying a mechanism that generates distinct emotions and other affective conditions seems not so hard—what is hard is to make it all believable. As I just indicated, a key issue is the need for affective artifacts to be able to parse the environment so as to understand its beneficial and harmful affordances—a crucial requirement for consistency, and thus also for believability. And a prerequisite for doing this is a coherent and relatively stable value system in terms of which the environment is appraised. As we indicated in OCC (and as illustrated in figure 6.1), such a system, at least in humans, is an amalgam of a goal hierarchy in which at least some of the higher-level goals are sufficiently enduring that

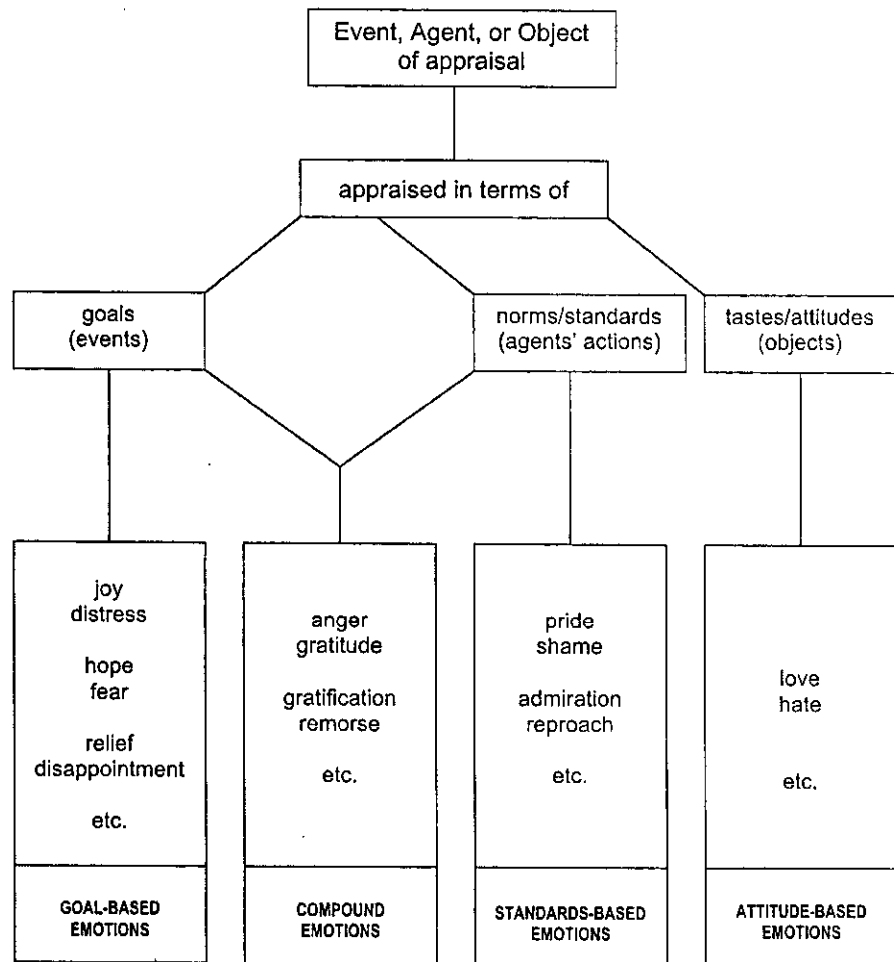


Figure 6.1 The relation between different things being appraised, the representations in terms of which they are appraised, and the different classes of resulting emotions.

they influence behavior and emotions over an extended period (rather than transiently), a set of norms, standards, and values that underlie judgments of appropriateness, fairness, morality, and so on, and tastes and preferences whence especially value-laden sensory stimuli acquire their value.

Another respect in which emotional reactions and their concomitant behaviors need some degree of consistency has to do with emotion intensity. It is not sufficient that similar situations tend to elicit similar emotions within an individual. Similar situations also elicit emotions of comparable intensity. In general, other things (external circumstances, and internal conditions such as moods, current concerns, etc.) being equal, the emotions that individuals experience in response to similar situations, and the intensity with

which they experience them, are reasonably consistent. Emotionally volatile people explode with the slightest provocation while their placid counterparts remain unmoved. In this connection, I'm reminded of a colleague (call him G) whom my (other) colleagues and I know to be unusually "laid back" and unemotional. One day several of us were having lunch together in an Italian restaurant when G managed to splash a large amount of tomato sauce all over his brilliant white, freshly laundered shirt. Many people would have become very angry at such an incident—I for example, would no doubt have sworn profusely, and for a long time! G, on the other hand, said nothing; he revealed no emotion at all—not even as much as a mild kind of "oh dear, what a bother" reaction; he just quietly dipped his napkin into his water and started trying to wipe the brilliant red mess off his shirt (in fact making it worse with every wipe), while carrying on the conversation as though nothing had happened. Yet, unusual as his nonreaction might have been for people in general, those of us who witnessed this were not at all surprised by G's reaction (although we were thoroughly amused) because we all know G to be a person who, when he emotes at all, consistently does so with very low intensity—that's just the kind of person he is, that's his personality.

6.3 Consistency and Variability in Emotion-Related Response Tendencies

The tomato sauce episode not only highlights questions about emotion intensity, it also, for the same reason, brings to the fore the question of the relation between (internal) emotional states and their related behaviors. To design a computational artifact that exhibits a broad range of believable emotional behavior, we have to be able to identify the general principles governing the relation between emotions and behavior, or, more accurately, behavioral inclinations, because, as Ekman (e.g., 1982) has argued so persuasively, at least in humans, social and cultural norms (display rules) often interfere with the "natural" expression (both in the face, and in behavior) of emotions.

Associated with each emotion type is a wide variety of reactions, behaviors, and behavioral inclinations, which, for simplicity of exposition, I shall refer to collectively as "response tendencies" (as distinct from responses). Response tendencies range from involuntary expressive manifestations, many (e.g., flushing) having immediate physiological causes, through changes in the way in which

information is attended to and processed, to coping responses such as goal-oriented, planned actions (e.g., taking revenge). From this characterization alone, it is evident that one of the most salient aspects of emotional behavior is that some of it sometimes is voluntary and purposeful (goal-oriented, planned, and intentional) and some of it is sometimes involuntary and spontaneous—as when a person flies into an uncontrollable rage, trembles with fear, blushes with embarrassment, or cries with joy.

Figure 6.2 sketches a general way of thinking about the constraints on the response tendencies for emotions. It shows three major types of emotion response tendencies (labeled “expressive,” “information-processing,” and “coping”), each of which is elaborated below its corresponding box. The claim is that *all* emotion responses have these three kinds of tendencies associated with them. Note, however, that this is *not* the same as saying that in every case of every emotion, these tendencies have observable concomitants—they are *tendencies* to behave in certain ways, not actual behaviors. The first group—the *expressive* tendencies—are the usually spontaneous, involuntary manifestations of emotions that are often referred to by emotion theorists (following Darwin) as emotional expressions. These expressive tendencies are of three kinds: *somatic* (i.e., bodily), *behavioral*, and *communicative* (both verbal and nonverbal). Consider first the somatic tendencies. These are almost completely beyond the control of the person experiencing the emotion. For instance, the box marked “somatic” in figure 6.2 has a parenthetical “flushing” in it. This (and the other parenthetical entries) is presented (only) as an example of the kind of response tendencies that one might expect to find in the case of anger; it should be interpreted as indicating that when someone is angry, one possible somatic manifestation is that the person grows red in the face. Notice that this is not something that he or she chooses to do. We do not choose to flush—our physiology does it for us, without us asking.

The next class of expressive tendencies are the behavioral ones. Again, these tendencies are fairly automatic, often hardwired, and relatively difficult (although not always impossible) to control; they are spontaneous actions that are rarely truly instrumental (although they might have vestigial instrumentality), such as kicking something in anger. So, to continue with the example of anger, I have in mind not the reasoned planful behaviors that might be entertained as part of a revenge strategy (they belong to the

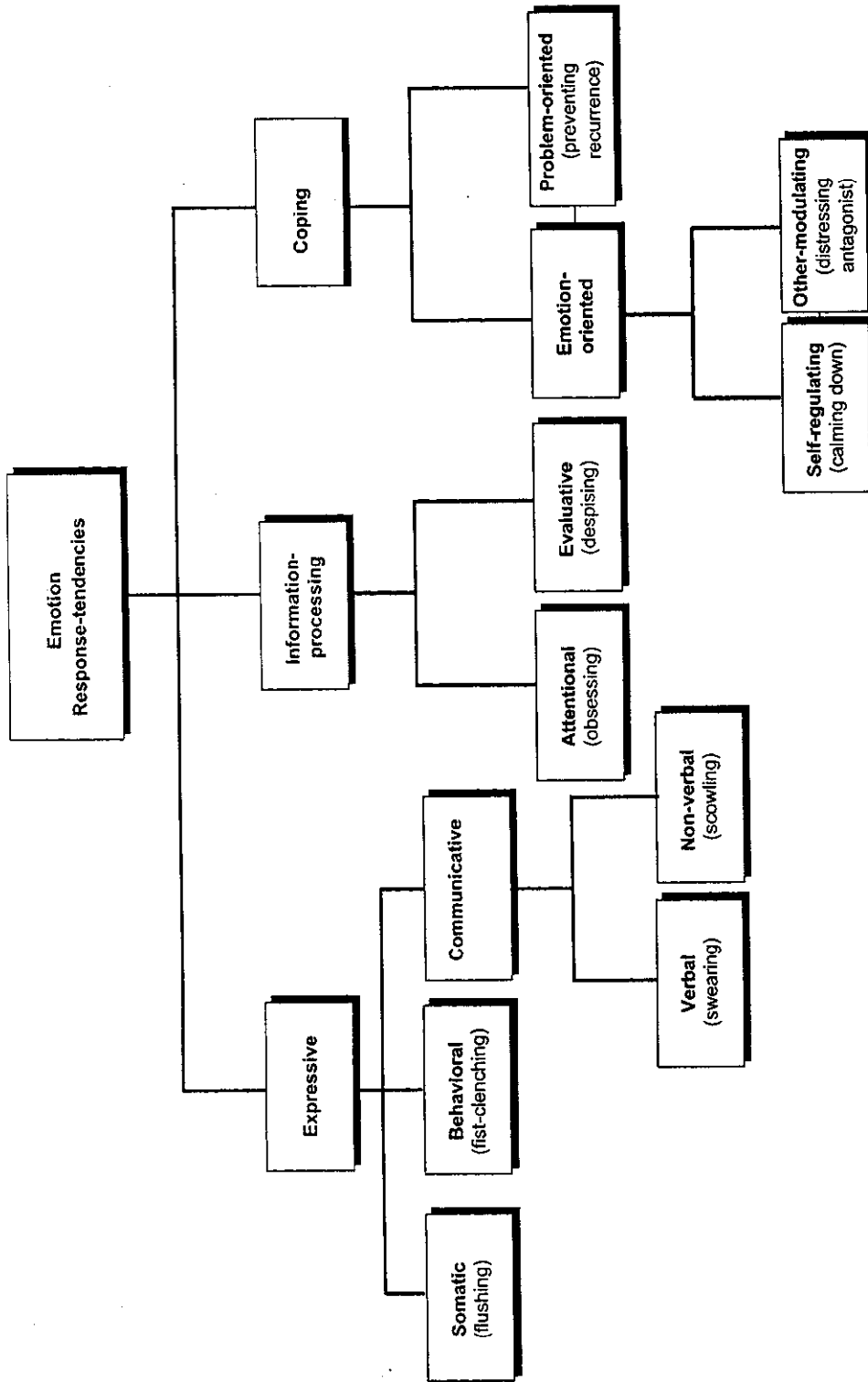


Figure 6.2 Proposed analysis of the behavioral structure of emotions. The parenthetical entries in the leaf nodes are intended as examples of the different kinds of response tendencies, in this case instances of response tendencies that might be associated with anger are indicated.

“coping” category), but the more spontaneous tendencies to exaggerate actions (as when one slams a door that one might have otherwise closed quietly), or the tendency to perform almost symbolic gestural actions (albeit, often culturally learned ones) such as clenching one’s fist.

Finally, I have separated out communicative tendencies (while realizing that symbolic acts such as fist clenching also have communicative value) as a third kind of expressive response tendency. Still, I wish here to focus more on communication through the face, because historically this has been so central to emotion research. Communicative response tendencies are those that have the capacity to communicate information to others, even though they are often not intended to do so. They have communicative value because they are (sometimes pan-culturally) recognized as *symptoms* of emotions. They include *nonverbal* manifestations in the face, including those usually referred to by emotion theorists as “facial expressions” (e.g., scowling, frowning of the brow), as well as *verbal* manifestations (e.g., swearing, unleashing torrents of invectives), and other kinds of oral (but nonverbal) responses such as growling, screaming, and laughing.

The second, *information processing*, component has to do with changes in the way in which information is processed. A major aspect of this is the diversion of *attention* (again often quite involuntary) from those tasks that were commanding resources prior to the emotion-inducing event to issues related to the emotion-inducing event. One of the most striking cases of the diversion of attentional resources is the all-consuming obsessive focus that people often devote to situations that are powerfully emotional. In humans, this obsessive rumination can be truly extraordinary and often quite debilitating, as so convincingly depicted in much of the world’s great literature—consider, for example, Shakespeare’s *Othello*. The second part of the information processing response has to do with updating beliefs, attitudes, and more generally *evaluations* about other agents and objects pertinent to the emotion-inducing event—you increasingly dislike your car when it repeatedly infuriates you by breaking down on the highway, whereas your liking for an individual increases as he or she repeatedly generates positive affect in you (Ortony 1991).

Finally, there are coping strategies, of which I have identified two kinds. One of these, *problem-oriented coping*, is what emotion theorists usually have in mind when they talk about coping;

namely, efforts to bring the situation under control—to change or perpetuate it—with the goal of improving a bad situation, or prolonging or taking advantage of a good one. In the case of anger, people often seek to do something that they think might prevent a recurrence of the problem, or that might somehow fix the problem.

The more interesting kind of coping is *emotion-oriented coping*. This kind of coping has to do with managing emotions themselves—either one's own, or those of some other agent or agents involved in the emotion-inducing situation. *Self-regulating* emotion-oriented coping responses focus on one's own emotions. For example, if I am angry I might try to calm down, perhaps as a precursor to developing a sensible plan to solve the problem, or perhaps simply because I don't like the feeling of being out of control. The *other-modulating* emotion management strategies can serve various purposes. For instance, if I induce distress in you because of what you did to me, not only might it make me feel better (i.e., it might help me to manage my own emotion of anger, hence the association in the figure between self-regulating and other-modulating responses), but it might also make you more likely to fix the problem you caused me (hence the link between emotion-oriented and problem-oriented responses). So, for example, suppose you are angry at somebody for smashing into your car. Developing or executing a plan to have the car fixed is a problem-oriented response, as would be a desire to prevent, block, or otherwise interfere with the antagonist's prospects for doing the same kind of thing again. But one might also try to modulate the antagonist's emotions by retaliating and getting one's own back so as to "make him pay for what he did to me," or one might try to induce fear or shame in him to make him feel bad, all with a view to making one's self feel better. There is no requirement that any of these responses be "rational." Indeed, if we designed only rational emotion response-tendencies into our emotional agents, we would almost certainly fail to make our emotional agents believable.

So the general claim is that a major source of consistency derives from the fact that all emotions constrain their associated response tendencies and all emotions have all or most of these tendencies. It should be clear from this discussion, and from table 6.2 (which indicates how the various constraints might be manifested in the emotions of anger and fear) that there is plenty of room for individual variation. Just as in the case of the emotions themselves, much of this variation is captured by traits—so many, although

Table 6.2 Sample manifestation of the different components for fear emotions (upper panel) and for anger emotions (lower panel)

Expressive	Somatic	Trembling, shivering, turning-pale, piloerection
	Behavioral	Freezing, cowering
	Communicative nonverbal	Screaming
Information Processing	Attentional	Obsessing about event, etc.
	Evaluative	Disliking source, viewing self as powerless/victim
Coping	Emotion Self-regulating	Calming down, getting a grip
	Emotion Other-modulating	Scaring away
	Problem-oriented coping	Getting help/protection, escaping, eliminating threat
Expressive	Somatic	Shaking, flushing
	Behavioral	Fist-clenching
	Communicative verbal	Swearing
	Communicative nonverbal	Scowling, frowning, stomping, fist-pounding, etc.
Information Processing	Attentional	Obsessing about event, etc.
	Evaluative	Disliking and despising source
Coping	Emotion Self-regulating	Calming down, getting a grip
	Emotion Other-modulating	Causing distress to antagonist
	Problem-oriented coping	Preventing continuation or recurrence of problem

not all, of the ways in which a timid person responds to anger-inducing situations are predictably different from the ways in which an aggressive person responds.

6.4 Why Personality?

Traits are the stuff of personality theory. Personality psychologists disagree as to whether personality should be viewed merely as an empirical description of observed regularities, or whether it should be viewed as a driver of behavior. But for people interested in building affective artifacts, personality can only be interesting and relevant if one adopts the second position. If one really wants to build believable emotional agents, one is going to need to ensure situationally and individually appropriate internal responses (emotions), ensure situationally and individually appropriate external responses (behaviors and behavioral inclinations), and arrange for sensible coordination between internal and external responses. Situationally appropriate responses are controlled by incorporating models of emotion elicitation and of emotion to emotion-responses of the kind I have just outlined. But to arrange

for individual appropriateness, we will have to incorporate personality, not to be cute, but as a generative engine that contributes to coherence, consistency, and predictability in emotional reactions and responses. The question is, how can we incorporate personality into an artifact without doing it trait by trait for the thousands of traits that make up a personality? In their famous 1938 monograph, *Trait Names: A Psycho-lexical Study*, Allport and Odbert identified some 18,000 English words as trait descriptors, and even though many of the terms they identified do not in fact refer to traits, the number still remains very large.

Trying to construct personalities in a more or less piecemeal fashion, trait by trait, is probably quite effective if the number of traits implemented is relatively small and if the system complexity is relatively limited. To some extent, this appears to be the way in which emotional behaviors and expressions are constrained in Cybercafé—part of Hayes-Roth's Virtual Theater Project at Stanford (e.g., Rousseau 1996), and to an even greater extent, in Virtual Petz and Babyz (see Stern, chapter 12 of this volume), and anyone who has interacted with these characters knows how compelling they are. However, if one has more stringent criteria for believability—as one might have, for example, in a soft-skills business training simulation, where the diversity and complexity of trait and trait constellations might have to be much greater—I suspect that a more principled mechanism is going to be necessary to produce consistent and coherent (i.e., believable) characters. Note, incidentally, that this implies that “believability” is a context-, or rather application-dependent notion. A character that is believable in an entertainment application might not be believable in an education or training application.

One solution to the problem of how to achieve this higher level of believability is to exploit the fact that traits don't live in isolation. If we know that someone is friendly we know that he has a general tendency or disposition to be friendly relative to people in general; we know that in a situation that might lead him to be somewhere on the unfriendly-friendly continuum, he is more likely to be toward the friendly end. However, we also know some other very important things—specifically, we know that he is likely to be kind, generous, outgoing, warm, sincere, helpful, and so on. In other words, we expect such a person to exhibit a number of correlated traits. This brings us back to the question of behavioral coherence. There is much empirical evidence that traits clus-

ter together and that trait space can be characterized in terms of a small number of factors—varying in number from two to five, depending on how one decides to group them. For our purposes here, the question of which version of the factor structure of personality one adopts may not be crucial (although I do have a personal preference). What matters is that the factor structure of trait space provides a meaningful way to organize traits. What matters is that it provides a meaningful and powerful reduction of data to note that people whom we would normally describe as being outgoing or *extroverted* (as opposed to introverted) tend to be sociable, warm, and talkative, and that people who are forgiving, good-natured, and softhearted we generally think of as *agreeable* or likeable (as opposed to antagonistic). Similarly, people who are careful, well organized, hard working, and reliable we tend to think of as being *conscientious* (as opposed to negligent). These (extroversion, agreeableness, and conscientiousness) are three of the “big five” (e.g., McCrae and Costa 1987) dimensions of personality—the other two being *openness* (as opposed to closed to new experiences), and *neuroticism* (as opposed to emotional stability).

The key point here is that such clusters, such groups of tendencies to behave and feel in certain kinds of ways, constitute one source of behavioral and emotional consistency and hence predictability of individuals. Viewed in this way, personality is the engine of behavior. You tend to react this way in situations of this kind *because* you are that kind of person. Personality is a (partial) *determiner* of, not merely a summary statement of, behavior. Consistent with this view (which is certainly not shared by all personality theorists) is the fact that some components of personality appear to be genetically based. All this suggests that to build truly believable emotional agents, we need to endow them with personalities that serve as engines of consistency and coherence rather than simply pulling small groups of traits out of the thin air of intuition.

A general approach to doing this would be to identify generative mechanisms that might have the power to spawn a variety of particular states and behaviors simply by varying a few parameters. Many of the proposals in the personality literature provide a basis for this kind of approach. For example, one might start with the distinction between two kinds of regulatory focus (e.g., Higgins 1998), namely, *promotion* focus in which agents are more

concerned with attempting to achieve things that they need and want (e.g., they strive for nurturance, or the maintenance of ideals). Promotion focus is characterized as a preference for gain-nongain situations. In contrast, with *prevention* focus, agents seek to guard against harm (e.g., they strive for security) and exhibit a preference for nonloss-loss situations. Thus regulatory focus is a fundamental variable that characterizes preferred styles of interacting with the world. Different people at different times prefer to focus on the achievement of pleasurable outcomes (promotion focus), or on the avoidance of painful outcomes (prevention focus). These are essentially the same constructs as approach motivation and avoidance motivation (e.g., Reville, Anderson, and Humphreys 1987), and are closely related to the idea that individuals differ in their sensitivity to cues for reward and punishment (Gray 1994). This can be clearly seen when we consider people's gambling or sexual behavior (sometimes there's not much difference): Those who are predominantly promotion focused (sensitive to cues for reward) focus on the possible gains rather than the possible losses—they tend to be high (as opposed to low) on the personality dimension of impulsivity; those with a prevention focus (sensitive to cues for punishment) prefer not to gamble so as to avoid the possible losses—these people tend to be high as opposed to low on the anxiety dimension.

If an individual prefers one regulatory strategy over another, this will be evident in his behaviors, in his styles of interaction with the world, and with other agents in the world, and as such, it constitutes one aspect of personality. Probably the most productive way to think about regulatory focus is that in many of our encounters with the world, a little of each is present—the question then becomes, which one dominates, and to what degree. Different people will have different degrees of each, leading to different styles of interacting with the world. Still, some of each is what we would ordinarily strive for in designing an affective artifact. Without some counterbalancing force, each is dysfunctional. For example, unbridled promotion focus is associated with a high tolerance for negativity (including a high threshold for fear, pain, and the like), and that comes pretty close to being pathologically reckless behavior.

I think it is possible to exploit these kinds of ideas in a principled way in designing our artifacts. We might start with the ideas of psychologists such as Eysenck, Gray, Reville, and others (e.g., Rolls 1999; chapter 2 of this volume) who take the position that

there are biological substrates of personality (such as cue sensitivity). The virtue of this kind of approach is that it provides a biological basis for patterns of behaviors and, correspondingly, emotions, which can serve as the basis for generating some sort of systematicity and hence plausibility or believability of an artificial agent. Which particular activities a human agent actually pursues in the real world is of course also dependent on the particular situation and local concerns of that agent, as well, no doubt, as on other biologically based determinants of other components of personality. But at least we have a scientifically plausible and computationally tractable place to start, even though the specification of exactly how this can be done remains a topic for future research.

Discussion: On Modeling Emotion and Personality

Bellman: So you are telling me that personality would be this core biological basis that somehow constrains behavioral inclinations. I have been bothered for a long time about a lot of research in emotions, because I am confused by the tendency to want to have oversimplified models—why people keep trying to reduce the space to two or three bases. There are many disciplines that I have been in which try to take complex multidimensional problems and reduce them to two or three bases. And, yes, you can do that at some level; but you usually lose most of the interesting stuff when you do that.

Sloman: I have a worse problem: Why try to find any number of dimensions as opposed to finding what the underlying architecture is and generating these things?

Bellman: But the underlying architecture doesn't have to be something with only two or three reinforcement/nonreinforcement bases. Why should that be the underlying architecture?

Rolls: The only theory is that one tries to get a principled approach here instead of doing something like factor analysis. The idea is to say something like this: So, what is it that causes emotion? If one would recreate and operationalize things where reinforcers are involved—most people find it difficult to think of exceptions to that—then one ought to pursue that idea and ask: What sorts of reinforcers are there? You know, you can give positive and negative reinforcement, you can withdraw positive and negative reinforcement. The second question is: What comes out of that? The

nice thing that Andrew is pointing us towards here is that personality might drop out of that research. If some individuals, by their genes, were a bit less sensitive to nonreward, or a bit less sensitive to punishment, it turns out that you would categorize them as extraverts. And so, one gets the dimension of personality without having to buy into any sort of special engineering specifications.

Ortony: But personality has consequences, because it constrains behaviors down the road, individual behaviors and motivations, and it makes one more likely to gamble in casinos and more likely to engage in unsafe sex and more likely to do a whole host of things which actually make people look as though they are individuals with some sort of stable underlying behavioral patterns.

Rolls: The idea is that here, then, is a sort of scientifically principled way to get personality. And I agree with your notion about consistency, but it is one of the quite nice things that come with personality. Notice that consistency is slightly different to persistence of emotions: If you have a nonreward, it's helpful on the short-time scale to keep your behavior directed towards a reward. So, the emotional state ought to be continuing slightly. Consistency, on the other hand, is more of a long-term requirement for believable agents: Next time you come round to that similar situation for that individual, it makes sense if they behave in a somewhat similar way. At least we as biological organisms do, perhaps for the reasons that we have discussed.

Bellman: The comment was not that there aren't some important principles. It is exactly how we model those important principles. I will give you a simple example: If we take language generation, we can model it, as we have learned to do over time. It has been very difficult, with all sorts and kinds of generative grammars. That's a very different kind of modeling than from a simple combinatorics. I don't know any cases, but one could imagine language as having been modeled as if it were a simple combinatoric problem. And eventually, people shifted to more interesting underlying modeling with grammars. That's part of my point: I don't see any reason why we should suppose, just because of a positive modeling, a simple combinatoric space. That's what my comment was directed towards, not the lack of principles.

Slooman: But are you talking about building synthetic artifacts for some purpose, which may be useful, or are you talking about how human beings work? Because, if it's the latter, there are going to be

constraints, and you can go and find out the biological bases, you can go and find out how the brain is involved.

Bellman: But there are lots of constraints that we know about in human language generation.

Sloman: So, the answer to your question is: Technically, you can take as many constraints as you get a handle on. And you do the best you can. And then someone else may come and have a better solution.

Bellman: Ok. I am suggesting a different style of modeling for it. I think that there is a long history of emotional modeling.

Petta: If you talk about human beings, the society and the environment as such are so complex that perhaps you have to leave out that part and just concentrate on the individual. You make an analysis of how the individual describes itself and end up with this collection of traits, which always just refers to the first person point of view. What I think is important however, especially when you are talking about artifacts, is that perhaps way more efforts should be put into performing a thorough lifeworld analysis—to use the terminology of Agre.¹ You would have to look at the whole system, what the environment provides and what kinds of couplings there are between a single architecture and between different instances of the architecture and/or what could happen in the environment, what kinds of dimensions, effects, and kinds of dynamics are relevant for your artifact. Personality, after all, is also something that is perceived about the other.

Sloman: Could you give an example of the kind of coupling you talked about?

Petta: Especially when you design an artifact, you want it to behave within given constraints. These constraints typically are characterized by, for instance, what kinds of interactions you want to occur, what kind of dynamics, how you want to stabilize its behavior. And once you know that, you can take a look at what happens in that environment, what can take place over a certain amount of time and how that relates to the possibilities of the agent, what its perceptual capabilities are, what its choice of actions is. Then you can start to consider how to constrain or direct those elements. Where do you put the incentives, where do you rather expand, where do you rather suppress? And these, in some, turn out to be biases, constraints, which again can have their own dynamics coupled to the environment, and which, I would

assume, should lead to something recognizable as a certain personality.

Ortony: Part of this, I think, has to do with a fact that I did not talk about, the appraisal side. Also, you want a level of description, if you are thinking about this in general, that goes beyond any particular design intention we may have. So, in thinking in general about what you want to do when you construct believable agents, you are not going to have one set of criteria for a person or a system, and a different set of criteria for what makes it believable. For a pet robot, for example, you have to have some level of description that characterizes the interaction which organisms are likely to have in a physical world. So, this gets you to things like this, but stuff happens that gets appraised. Again, there are individual differences, but there are also similarities with respect to our goals. The norms or standards that we use to make judgments of the kind that lead us to approve or disapprove of things, which is different from goals, although they could in fact be collapsed if you said that you had a goal to maintain order or something. But the point is, it does not matter what the goals of the organism are. It only matters that it indeed *has* some goals. It is very difficult to imagine building an emotionally believable artifact which didn't have goals.

So, once you admit that it has got goals, the architecture should be such that it's impervious to the particular kind of goals. It only cares what happens when goals are satisfied or blocked or failed. This really comes down to how you characterize the underlying value system in terms of which appraisals of the environment you are constructing. Is that an answer to your question or not?

Petta: Partly, because, actually, I refer to the box with norms and standards (cf. figure 6.1), because this is where you introduce aspects of the society which are beyond the individual. So, this is just one very evident spot where you gather stuff that is external to the single individual and put it into your picture.

Slovan: But if that does not get internalized by the individual, it has no effect on the individual's behavior.

Petta: Yes, sure. Obviously, there must be a connection, there must be a coupling.

Slovan: Can I, in this context, say something which, at first, will contradict with what you say? You have been saying that you need consistency because of the predictability of personality. Now, if you actually look at human beings, but not necessarily at other

animals, you will find that there may be consistency in a particular individual's behavior in a certain context only. If you put him in another context, you will get a different collection of behaviors, which is itself consistent. So, at home with your family, you may be kind and generous and thoughtful, whereas being in the aggressive MIT AI Lab or in the office, where there is a lot of competition and insecurity, you may suddenly behave in a very different way, but consistently in itself. I think that would not be possible for other animals.

Ortony: It might actually: animal as parent and animal as hunter, for example.

Sloman: Yes. In that case, it may be a very general characterization that, in some sense, there is not a single personality, there are sub-personalities which get turned on and off by the context.

Petta: By the context, by the environment. That's precisely my point.

Ortony: Yes and no, is my reaction. At some level, of course, it's true that we are going to behave less aggressively in an environment in which the interpersonal interactions are characterized by love, affection, and familial relations, than when we are in a hostile or in a competitive environment—at the workplace, for example.

Elliott: You can have one personality that appraises everything always the same way, but it's appraising different things at different times.

Sloman: But even cues for punishment or reward might be a variable factor. They may be consistently low in this situation and consistently high when you are in that situation.

Elliott: Also, you can't leave out that individuals are highly affected by moods, too, in that you can characterize being in the workplace as placing yourself in an aggressive mood.

Ortony: Let's take a dimension like friendliness—a person you characterize as a friendly person would be represented on one side of a curve for "friendliness" in a group. But, for this particular individual, it is still true that there is a distribution of behaviors relative to his own behaviors, such that some of his behaviors are friendly and some are unfriendly. But this distribution lies inside the "friendly" sector with respect to the reference group.

Sloman: And that distribution might shift with context for the same individual.

Ortony: Yes, it could shift, but it's still probably the case that a person whom you would regard as aggressive is going to have more aggressive behaviors in the aggressive roles.

Sloman: I think what you are saying is that there may be an even deeper consistency. I suspect that for some people, there is a degree of integration that is higher than for others. And in an extreme case you get personality disorders where this mechanism is going badly wrong.

Rolls: So, is the bottom line of this that your sensitivity to reward and punishment could be relatively set, but the actual coping strategies that you are bringing into play in different environments are appropriate for that particular environment? For example, if there is something you can do about it, you might be angry; but if there's nothing you can do about it, you might be sad? Is that one way to try to rescue a sort of more biological approach?

The basic biology might—successive to the reward and punishment—be unchanged, but then you have different, as it were, coping strategies.

Sloman: This is an empirical question, and I have no reason to think that, at least for humans, it's more like what I said than like what you said. But I could be wrong. We have to investigate.

Rolls: Yes, that's right. But I think it's worth underlining the fact that there are at least two possibilities to explain context dependency of personality.

Note

1. P. E. Agre, *Computation and Human Experience* (Cambridge University Press, Cambridge, 1997).

References

- Allport, G. W., and Odbert, H. S. (1936): Trait Names: A Psycho-Lexical Study. *Psychol. Monogr.* 47 (1): whole no. 211.
- Averill, J. R. (1982): *Anger and Aggression: An Essay on Emotion*. Springer, Berlin, Heidelberg, New York.
- Colby, K. M. (1981): Modeling a Paranoid Mind. *Behav. Brain Sci.* 4: 515–560.
- Darwin, C. (1872): *The Expression of Emotions in Man and Animals*. Murray, London.
- Ekman, P., ed. (1982): *Emotion in the Human Face*. Cambridge University Press, Cambridge.
- Gray, J. A. (1994): *Neuropsychology of Anxiety*. Oxford University Press, Oxford, London, New York.

- Higgins, E. T. (1998): Promotion and Prevention: Regulatory Focus as a Motivational Principle. In M. P. Zanna, ed., *Advances in Experimental Social Psychology*. Academic Press, New York.
- McCrae, R. R., and Costa, P. T. (1987): Validation of a Five-Factor Model of Personality across Instruments and Observers. *J. Pers. Soc. Psychol.* 52: 81-90.
- Ortony, A. (1991): Value and Emotion. In W. Kessen, A. Ortony, and F. Craik, eds., *Memories, Thoughts, and Emotions: Essays in Honor of George Mandler*. Laurence Erlbaum Associates, Hillsdale, N.J.
- Ortony, A., Clore, G. L., and Collins, A. (1988): *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.
- Ortony, A., Clore, G. L., and Foss, M. A. (1987): The Referential Structure of the Affective Lexicon. *Cogn. Sci.* 11: 341-364.
- Revelle, W., Anderson, K. J., and Humphreys, M. S. (1987): Empirical Tests and Theoretical Extensions of Arousal Based Theories of Personality. In J. Strelau and H. J. Eysenck, eds., *Personality Dimensions and Arousal*, 17-36. Plenum Press, London.
- Rolls, E. (1999): *The Brain and Emotion*. Oxford University Press, Oxford, London, New York.
- Roseman, I. J., Antoniou, A. A., and Jose, P. E. (1996): Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory. *Cogn. Emotion* 10: 241-277.
- Rousseau, D. (1996): Personality in Computer Characters. Paper Presented at the Annual Meeting of the American Association of Artificial Intelligence. In: H. Kitano, ed., "Entertainment and AI/A-Life", AAAI Workshop Technical Report WS-96-03, AAAI Press, Menlo Park, CA, pp. 38-43.
- Scherer, K. R. (1997): Profiles of Emotion-Antecedent Appraisal: Testing Theoretical Predictions Across Cultures. *Cogn. Emotion* 11: 113-150.