Modelling Strategic Emotions

William J. Turkel
Brain and Cognitive Sciences
Massachusetts Institute of Technology
E10-105B
Cambridge, MA 02139
wjt@psyche.mit.edu

I am interested in mathematical and computational modelling of emotions, and in testing theories of emotion experimentally. The perspective that I take is a strategic one, developed primarily in the economic literature. Briefly, theories of emotions tend to be of two sorts. Romantic theories stress the divide between rational thought, which originates in the mind and is the product of civilization, and the irrational emotions, which originate in the body and are a heritage of animal ancestors. Under such a view, the emotions are typically held to be a source of wisdom, creativity or authenticity. Such theories take many guises in psychology: id and superego, right and left brain, limbic system and cortex, and so on.

The strategic theory of emotion is distinctly unromantic. From the theory of evolution, we know that the behavior and other complex traits of organisms typically exist to promote the survival of the genes carried by the individuals that bear them. Natural selection works relentlessly against organisms that do not maximize the survival of their genes. A fairly straightforward application of this idea would seem to be that the most successful organisms will always act in self-interest. But, of course, this is not always the case. The strategic theory resolves the apparent contradiction in the following way: in many situations, the conscious pursuit of self-interest will not lead to its attainment. An act that appears to be irrational or to go against self-interest when considered in isolation, can be seen to be advantageous when its effects are averaged over many situations in the individual's lifetime.

Consider some simple examples. If you know that someone will respond to an affront with retaliation, even at a great cost to themselves, then you are less likely to mess with them. Surely, responding in such a way is not in that person's best interest when viewed with respect to a single, costly retaliatory act. But averaged over a longer timescale, that individual enjoys the benefit of having a reputation, and as a consequence actually has to retaliate less. The fact they cannot stop themselves from retaliating if provoked makes their threat to do so credible. In this way, a passion such as honor, or a thirst for vengence, acts as a guarantor for a threat of retaliation. Similarly, if you want someone to do something for you, and you promise to reciprocate, that person will be more likely to perform the action if they know that you are true to your word. Being true to your word means that you may have to do things that go against your self-interest in the future: perhaps following through on a promise that does you more harm than good.

Crucially, these kinds of problems arise only when dealing with other intelligent agents. Having a taste for fruit will ensure that you eat one when you encounter it; nothing more is necessary. In dealing with other agents, however, reason is not always sufficient to prevent being outsmarted. An individual that threatens to retaliate may be bluffing. Other individuals then need a way to detect bluffing. The retaliator, in turn, must find a way of indicating that they really are not bluffing—the threat must be credible.

I am currently working on mathematical models of strategic emotions, using a variety of simulation techniques (including agent-based models) to study scenarios under which such emotions might have evolved. One line of inquiry focuses on the role of communication and perception in signalling emotional commitment. In human interaction, facial expressions, posture and gesture, and visible autonomic responses are used to convey information about emotional state. Here, ideas from evolutionary game theory, linguistic pragmatics, and computational modelling of language evolution play a role. I am also interested in applying computational and statistical learning theory to the problem of explaining how emotional agents might learn to bargain with one another.

The strategic perspective has important implications for affective computing and artificial emotions. Rather than portraying emotions as random disruptive or associative mechanisms that foster creativity, the focus is on the way that emotions are integrated into a cognitive system that increases the agent's fitness. Arguably, an artificial system will need to be able to make effective threats and promises in order to do human things, like bargain. According to the strategic the-

ory, such threats must be credible. In human beings, emotions such as fear, anger and honor underwrite the credibility. But if a user has no reason to fear the computer, then its threats are empty. Likewise, if the computer has no power to fulfill promises, then it cannot make them. And a computer that cannot be angry with you cannot really like you either.