The Sciences of the Artificial Emotions: Commentary on Aylett & Paiva

Jonathan Gratch
University of Southern California

Recent years have seen the rise of a remarkable partnership between social and computational sciences around the phenomena of emotions. Rallying around the term *Affective Computing,* this research can be seen as revival of the cognitive science revolution of the '50s and '60s, albeit garbed in the cloak of affect, rather than cognition (e.g., see Picard, 1997; K. R. Scherer, Bänziger, & Roesch, 2010). Aylett and Paiva are clearly in the vanguard of this enterprise, but they should be seen in the context of many related efforts struggling to reconcile socio-emotional theories with the rigorous of computational modeling. In this commentary, I'd like to address this broader undertaking and will argue that the field has much to learn by revisiting the original motives and lessons behind the cognitive science revolution. Specifically, I will summarize the arguments made by Herb Simon in his seminal book *The Sciences of the Artificial* and argue for their continued relevance for both computational and psychological research on emotion.

In his 1969 book, Herb Simon argued that computational scientists bring a unique and complementary perspective to the challenge of understanding human intelligence. First, in contrast to the natural sciences which seek to describe intelligence as it is found in nature, the "artificial sciences" seek to describe intelligence as it "ought to be in order to *attain goals,* and to *function"* (italics his). This normative emphasize often leads to serviceable abstractions that crisply capture the essence of a phenomena while avoid the messy details inherent in how these functions are implemented in biological organisms. [i]

Second, computational scientists approach the problem of achieving these goals and functions with a mindset emphasizing *process.* Specifically, they conceptualize goal-directed behavior as an unfolding dynamic interaction between the intelligent artifact and its environment. Simon illustrated this point through his famous metaphor of an ant on the beach (pg. 53):

> He moves ahead, angles to the right to ease his climb up a steep dunelet, detours around a pebble, stops for a moment to exchange information with a compatriot. Thus he makes his weaving, halting way back to his home. … Viewed as a geometric figure, the ant's path is irregular, complex and hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant.

The insight here is that apparent complex behavior can often be reduced to simple goal-directed processes interacting over time with a complex environment.

Finally, the ultimate aim of a computational scientist is to produce a working artifact that realizes this design. Once produced, this computational model serves as an "empirical object" in ways that more conventional paper-and-pencil theories cannot. For example, a designer might attempt to characterize an ant's cognitive processes in terms of a minimal number of abstract goals and functions. By building a program that realizes these processes, and simulating the interaction of this model with complex

environments, the designer can empirically work out the implicit consequences of theoretical assumptions. And indeed, it is hard to imagine an "environment" more complex – and more in need of serviceable abstractions, complexity-reducing models, and tools for inferring far-reaching consequences – than the social landscape between two emotional people.

I'm concerned that research on computational models of emotion is losing sight of Simon's fundamental insight, and as a result, is missing an opportunity to bring simplicity and clarity to psychological theories. In a sense, computational researchers are taking psychological theory too literally and, by consequence, absorbing complexity akin to the path of Simon's ant into their models. Indeed, Simon's original argument was that psychological theories were sorely in need of the rational reinterpretation using the computational tools of function and process. Thus, by faithfully "re-implementing" a psychological theory, computational scientists are doing a disservice both to computational science and the original theory.

In the remainder of this commentary, I will argue this point with two examples of how some of the complexity of psychological models naturally belongs in the interaction between the organism and its environment, rather than within the structure of the theory. I will consider Scherer's sequential checking theory of emotion and argue that the sequential constraints are more naturally seen as arising from the interaction of a dynamic process with its environment. Next I will turn to representations of culture – the focus of Aylett and Paiva's work. Building on recent work by Yamagishi and colleagues (2008), I will argue that much of the complexity in modeling culture naturally resides in the environment, and not in the model as Aylett and Paiva suggest.

Most computational models of emotion draw inspiration from appraisal theory (see Marsella, Gratch, & Petta, 2010 for a recent review). As noted by Aylett and Paiva, one of the more recent and sophisticated versions of appraisal theory is the multi-level sequential checking theory of Scherer (Klaus R. Scherer, 2001). This process model of appraisal posits three levels of appraisal processing, innate (sensory-motor), learned (schema-based) and deliberative (conceptual) and posits a fixed sequential ordering of appraisals, based on physiological and functional considerations. Specifically, the model asserts that appraisals unfold in a sequential order beginning with relevance, implications for the self, coping potential and, finally, implications for other and social norms.

Stacy Marsella and I have argued elsewhere (Marsella & Gratch, 2009) that such complexity is unnecessary if one views appraisal as reflecting the unfolding interaction between an agent and its environment over time. Rather than reflecting hard cognitive constraints, the temporal sequencing of appraisals has more to do with the structure of the environment, the complexity of specific inferences that this environment demands (e.g., some appraisals of goal conduciveness are self-evident whereas others require complex inference or even planning), and what inferences are already active in working memory due to recent thoughts and events. In many cases, the sequence can unfold in the order Scherer suggests. For example, if an unexpected potential threat to self arises, it is natural to devote cognitive resources to appraise the consequences for the self before considering other social actors. On the other hand, if social concerns are central to the current environmental situation, the reverse pattern seems likely. For example, if we already suspect someone has violated a social norm, if they admit to

the crime it makes sense to jump straight to anger without waiting to calculate a cascade of irrelevant appraisals.

More generally, theories and models that posit appraisal as a separate process from cognition unnecessarily complicate the model by conflating *appraisal* and *inference*. Much of what passes as appraisal (e.g., calculations of likelihood, benefit and control) would be required by any goal-directed entity. From Simon's perspective, then, computational modelers should direct their efforts at finding which processes are minimally required to achieve goals, with the hope that apparently complex appraisal processes simply emerge from the behavior of these more basic systems interacting with the environment.

For another example, let's consider culture, the focus of Aylett and Paiva's article. Common sense and an abundance of research argue that people from different cultures behave differently and that some of this variance can be explained through broad cultural dimensions such as individualism and collectivism. The challenge for developers of computational models lies in where to attribute these differences. Cultural psychologists largely attribute these differences to internalized norms and values (e.g., Markus & Kitayama, 1991) – essentially locating the complexity of culture within the individual – and Aylett and Paiva are in good company in adopting this approach. More recent work, building on computational theory, points to a very different conclusion: that that the complexities implied by cultural variability more naturally reside in the environment (Adam, Shirako, & Maddux, 2010; Yamagishi, et al., 2008).

Using the computational framework of game theory, Yamagishi and colleagues reinterpret some of the key evidence for culture as a psychological construct by showing that cultural differences previously explained as culturally-varying internalized preferences are best seen as strategies adapted for variations in the local environment. For example, the classic study of Kim and Markus (1999) is oft-cited in defense of an internal individualistic vs. collectivistic difference between Americans and Asians. In this study, Kim and Markus showed that, when presented with an offer of a free pen, Americans would choose the least common color (indicating an internalized preference for individuality) whereas Japanese would chose the most common color (indicating an internalized preference for conformity). In a clever series of studies, Yamagishi et al. (2008) showed, in fact, that these preferences are dictated by the situation and not internalized norms. For example, when allowed to make the choice anonymously, Asians act identically to American participants. The implication is that, rather than being dictated by internalized preferences, behavior in this task is influenced by anticipated negative evaluations by other actors in the environment. This is not to say that all differences can be explained by differences in the environment, but, as Yamagishi concludes, "the field could benefit from examining what types of cultural differences are better explained and predicted through [examining difference in the environment], rather than the 'preference' approach frequently used in psychology" (Yamagishi, et al., 2008, p. 584).

Computational models can provide key insights into psychological processes and Aylett and Paiva's work is sure to open new avenues for research on emotion and culture. As this work proceeds, I've argued it is important to revisit Herb Simon's original insights into the value of a computational perspective. The human brain is awesome in its complexity and has stubbornly resisted naïve attempts by artificial

intelligence researchers to replicate its function. The response to these early setbacks has led some researchers of "the artificial" back to theories of human intelligence. The result is an exciting revival of interdisciplinary research, but oddly, some of this research has taken on (at least in my view) an unhealthy reverence for psychological theory. The computational sciences bring a unique perspective to the understanding of intelligence. Though our training in design and process, we more naturally approach complexity from the perspective of interactivity – i.e., that apparent complexity can arise from simple processes interacting with the environment. This wisdom is as true as when Simon noted over sixty years ago.

References

Adam, H., Shirako, A., & Maddux, W. W. (2010). Cultural Variance in the Interpersonal Effects of Anger in Negotiations. *Psychological Science, 21*(6), 882-889.

Kim, H., & Markus, H. R. (1999). Deviance or Uniqueness, Harmony or Conformity? A Cultural Analysis. *Journal of Personality and Social Psychology, 77*(4), 785-800.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224-253.

Marsella, S., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research, 10*(1), 70-90.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational Models of Emotion. In K. R. Scherer, T. Bänziger & E. Roesch (Eds.), *A blueprint for affective computing: A sourcebook and manual*. New York: Oxford University Press.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Scherer, K. R. (2001). Appraisal Considered as a Process of Multilevel Sequential Checking. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research* (pp. 92-120): Oxford University Press.

Scherer, K. R., Bänziger, T., & Roesch, E. (Eds.). (2010). *A Blueprint for Affective Computing: A sourcebook and manual*: Oxford University Press.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74*, 29-39.

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences Versus Strategies as Explanations for Culture-Specific Behavior. *Psychological Science, 19*(6), 579-584.

[i] The terms rational and normative have a complex and unfortunate relationship with emotion research. For historical reasons, the concept of rational behavior has often been held up in opposition to emotional behavior. In Simons terminology, rational simply refers to function. If we posit that emotions serve important intra- and inter-personal functions, then they are subject to the sort of rational analysis Simon proposes. Indeed, some of his work emphasized that emotion serves important cognitive functions that would be required of any intelligent entity, be it natural or artificial (Simon, 1967).