Adaptive **Behavior**

Forces, Fields, and the Role of Knowledge in Action: Commentary on Randall Beer "The Dynamics of Active Categorical Perception in an Evolved Model Agent"

Andy Clark
Cognitive Science Program, Indiana University

Beer's (2003) paper is a tour de force of detailed dynamical modeling, and provides a concrete sample of the kinds of understanding dynamicists may realistically hope to achieve. The analysis is thus, as Beer states, a "tool for building intuition", and in this it succeeds brilliantly. But it is also an attempt to show that dynamical approaches can get a foothold in the explanation of "minimally cognitive behaviors"; that is to say, behaviors that seem, on the surface at least, good contenders for more traditional forms of problem decomposition and analysis. In these brief comments, I want to focus on one important question that I think remains unanswered, and that bears rather directly on this enterprise of "scaling up".

The question concerns the notion of an integrated dynamical explanation itself. The point about such explanations, as I understand it, is to provide a kind of integrated window on the production of behavior. By an "integrated window" I mean a perspective that treats bodily, neural, and environmental factors and forces in a kind of common dynamical currency (a single mathematical language of trajectories, bifurcations, parameters, state variables, etc., and ultimately, perhaps, of differential or difference equations). In the opening comments on the more general notion of "situatedness", Beer suggests that "on this view, situated action is the fundamental concern and cognition is … one resource among many that can be brought to bear as an agent encounters its world". In the worked example of an agent that approaches circles and avoids diamonds, we see direct evidence of this in the claim that a certain three-dimensional projection provides a potent analytic tool. For this projection happens to be one that involves one environmental state variable (vertical object position), one body state variable (horizontal position relative to the object) and one neuronal state variable (output of interneuron 9). By dispensing with talk of representations and their contents, and restricting the depiction of the inner realm to a depiction of inner state alone (section 9.3), the dynamicist makes it easy to treat all three factors (bodily, neural and environmental) at once and on an even par, thus allowing projections based on any combinations that the theorist suspects might pay explanatory or predictive dividends.

But can we really *afford* to buy this flexibility by ignoring the apparently special role of some of these factors and forces in the production of intelligent behavior? Here is a very simple example of what I

Correspondence to: A. Clark, Cognitive Science Program, Indiana University, USA.
E-mail:
Tel.: +1-000-000, *Fax*: +1-000-000.

have in mind. Consider the account of approaching the circles and avoiding the diamonds, but now imagine that the catching agent is a human child, and the set-up some kind of video game. We know (for example, from the powerful dynamical analysis of the A-not-B error by Thelen, Schoner, Scheier, & Smith (2001) that there are many delicate balances that together determine on-the-spot action, such as reaching to one location rather than another. But we also know, from out own experience, that we can very often make a reach and know, while we do so, that we have made a mistake. In these cases, what we know seems, in a totally non-mysterious way, to outrun our bodily control. (Think of the sign on the faulty toilet that says Do Not Flush: we read the sign, understand it, and find ourselves flushing despite our best intentions.) The human child, playing the diamond/circle video game may surely have a similar experience at times: she will make a move that she knows, right away, will lead to a mistake, yet be unable to correct the error. My question is, how do we do justice to this kind of case?

Such mundane cases of self-conscious error make it seem as if, prima facie, there exist importantly distinct strands in the interlocking chains that lead to the production of action. Some of these chains seem to have more to do with bodily dynamics, habituation, and long-term learning, while others depend more on short-term states of information-based control. Since these states and processes constantly interact in the generation of behavior (including verbal behavior), we really do need the kind of integrative framework Beer and others propose. But the suspicion remains that the strands are importantly distinct. Surely it matters that, in the case of the mistaken flush, we knew we were making a mistake? One way to do justice to such an intuition is, of course, to depict at least one of the interlocking strands leading to action as involving a mental representation of the goal of not flushing, or of moving towards the circle, or of reaching for such-and-such a location. That such a strand failed to win the day does not make it unreal, nor does it blur (rather, it underlines) the difference between some factors and forces and others. All this relates, I believe, to the fact that the simple model agent makes use of a unitary neural resource for judgment and action, whereas in the human case, we seem to be deploying multiple specialized-yet-densely-interacting subsystems, some of which seem more concerned with fluent action-control and others more concerned with planning, judgment and categorization (see Milner & Goodale (1995), and discussion in Clark (1999)).

The account (section 5.3) of when the diamond/circle avoid/approach decision (in the simple model agent) is made is somewhat odd for a closely related reason. It purports to be probing the question of when the agent makes the decision to catch or avoid an object, and suggests that it is not until very near the moment of action that a decision has really been made. But this is because the criterion of decision is something like 'irreversible commitment to a specific response'. But notice that in the thicker world of human thoughts and reasons, we often seem to make firm decisions that alter over time (even without new input). I am not convinced that, in these thicker cases at least, it is somehow more correct to say that no real decision is made at those points. It seems perfectly fine to imagine both a firm decision and a subsequent change (even a subsequent unprompted change). Once more, it seems to me that the model agent does not yet show enough of a gap between behavior and considered judgment. We have to take its non-verbal behaviors or behavioral trajectories as exhaustive of its judgings, and this is a sign that there is something important, and perhaps (just perhaps) qualitatively different that still falls outside the scope of the model. (It would be natural to suspect, though I shall not pursue this here, that this something has much to do with our abilities to vehicle our thoughts in language, thus making them objects for our own attention and processing (see, for example, Dennett (1987) and Clark (1998)).

Notice that I am not suggesting that bodily or environmental factors cannot play, or help to play, the "special" kind of role at which I am trying to gesture. It is not, for example, that the neural strand is itself special. Rather, my worry is that *cognition* is special, and this may be so even if cognitive work can indeed be done by many means. Perhaps, in some advanced cases, we represent a goal by some canny mixture of environmental, neural, and bodily tinkering. We may, for example, represent some specific numerical goal or fact by using our fingers, or pen and paper. In such a case, as the brain–body–world system works to solve the problem, a wide variety of other factors and forces may yet intrude causing our actual behavior to drift from our target, just as it did in the case of the mistimed flush. With the cognitive action spread across brain, body, and world, we may still watch in horror as, despite our best intentions, we do the wrong thing.

My question is thus whether by eliding the difference between those factors and forces that affect action simpliciter and factors and forces (whether bodily, neural, or environmental) that affect action *by affecting what we think, judge, remember, and believe*, we might be suppressing important structure, and missing something that really matters. In his own description of the circles/diamonds agent, Beer notes that "In a very real sense, the evolved CTRNN [continuous-time recurrent neural network] does not 'know' the difference between circles and diamonds. It is only when embodied in its particular body and situated within the environment in which it evolved that this distinction arises over time through the interaction of these subsystems" (Beer, 2003: 000). Another way to raise the question I want to press is thus: what about an agent that *does* know the difference, yet still may act inappropriately? Can we afford, in attempting to understand this kind of more complex agent, to treat all the factors and forces interchangeably? If (as I suspect) we cannot, then is this a reason to once again pursue the use of content-ascribing glosses as a means of highlighting the special roles of specific elements in the dynamical mix? Is it really likely that, once we confront systems that *really do know the difference* between circles and diamonds, we will still fail to unearth inner states or processes that seem to code for the presence or absence of the features (in this case visually perceived squareness or circleness) in question?

As a kind of aside, it is possible that the debate concerning the need for a robust notion of misrepresentation (see the discussion in section 9.3) would also benefit from considering the class of cases considered earlier. For by respecting the gap between behavior and judgment, we drive a wedge between behavioral success or failure and how the agent represents the world to be. We thus make room for a notion of misrepresentation that is indeed different from simple failure of adaptive response.

Finally, I am aware that in pushing these issues, I may seem once more to be moving the goalposts: first from situated response to response in 'representation-hungry' cases, and now to something like 'making room for self-diagnosable action-judgment mismatches'. I do believe, though, that it is only by continuing to raise, in as straightforward a form as possible, the very hardest problems that we will get a sense of the ultimate power and scope of these new and exciting ways of thinking about mind, cognition and action.

References

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11*(4), 000–000.

Clark, A. (1998). Magic words: How language augments human computation. In S. Boucher and P. Carruthers (Eds.), *Thought and language* (pp. 162–183). Cambridge, UK: Cambridge University Press.

Clark, A (1999). Visual awareness and visuomotor action. *Journal of Consciousness Studies*, *6*(11-1), 1–18.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Pres).

Milner, A. & Goodale, M. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.

Thelen, E., Schoner, G., Scheier, C., & Smith, L. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences, 24*(1), 1–55.