Structured Representations in the Control of Behavior Cannot Be So Easily Dismissed: A Reply to Botvinick and Plaut (2006)

Richard P. Cooper
Birkbeck, University of London and University College London

Tim Shallice
University College London and Scuola Internazionale Superiore di Studi Avanzati

M. Botvinick and D. C. Plaut (2006) argued that many of the criticisms of their earlier simple recurrent network (SRN) model of routine sequential action raised by R. P. Cooper and T. Shallice (2006) were criticisms of the specific implementation rather than criticisms of the underlying theory. Cooper and Shallice (2006) reject this assessment and raise concerns with several implementational adjustments that Botvinick and Plaut made to address their criticisms of the SRN account. Moreover, Botvinick and Plaut are questioned for not addressing potential interactions between their suggested implementational changes. Cooper and Shallice also reconsider the implications of the role of the training set in shaping the SRN model's normal and error-prone behavior, the role of goals in their original interactive activation network model and routine behavior more generally, and the relation between the putative routine and nonroutine action control systems within the 2 models.

Keywords: routine action, action schema, goals, simple recurrent network, interactive activation network

Botvinick and Plaut (2006) essentially rejected our criticisms of their article on the simple recurrent network (SRN) model of action control as based on a series of inadequate characterizations of their model and unduly concrete interpretations of its potential. They claimed that several of our criticisms of their account of routine sequential behavior are criticisms of a specific implementation (the SRN model) rather than of their basic theoretical position. Instead, we continue to see fundamental problems in their position. We consider briefly their response to six of our specific observations and also to three of the wider issues we raised.

Specific Observations

Object Substitution Errors

In response to our concerns about the SRN model's account of object substitution errors, Botvinick and Plaut (2006) put forward a novel explanation of how such errors might arise. According to this view the errors are not generated by the SRN mechanism, but instead it is suggested that the fixate actions that are output by the SRN model might be error prone. This new explanation provides the SRN model with an additional source of error, orthogonal to the addition of noise to the context layer. It would be interesting to investigate the extent to which these different sources of error might correspond to different patient groups.

Richard P. Cooper, School of Psychology, Birkbeck, University of London, London, England, and Institute of Cognitive Neuroscience, University College London, London, England; Tim Shallice, Institute of Cognitive Neuroscience, University College London, and Cognitive Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy.

Correspondence concerning this article should be addressed to Richard P. Cooper, School of Psychology, Birkbeck, University of London, Malet Street, London WC1E 7HX, United Kingdom. E-mail: r.cooper@bbk.ac.uk

We have two concerns with the new account. First, as Botvinick and Plaut (2006) noted, including in the training corpus occasions on which malfunctions of the fixation mechanism occur will work against the production of object substitution errors, assuming the model can still learn adequately. The trained model will tend to automatically correct erroneous fixate operations by repeatedly fixating until the correct object is located. However, it is by no means obvious that learning will still be successful. If a fixate action results in fixating on an incorrect object, the SRN must learn to repeat the previous fixate action, effectively without modifying its context units. This revision is therefore likely to result in a model that weights more heavily its perceptual input following any fixate action and less heavily its context representations. This will make it less able to learn long-distance dependencies such as *if sugar has been added before cream then following the cream subsequence the model should move on to the drink sequence*; these are encoded in context representations.

An even more critical problem for this new account of object substitution errors concerns what happens following such an error. When we ran the SRN model past the point of error in the coffee-pouring situation described by Botvinick and Plaut (2006), it produced various behaviors. However, it never performed the rest of the coffee task without error. Typically the model floundered for some steps until it found its way into some point of a sequential attractor of a known task. Consider the SRN model's behavior following an error on the task-initial fixate action such that when making coffee it mistakenly begins by locating the teabag. In this case, the model's behavior consists of flawless preparation of tea, not coffee. In other words, object substitution errors on this account dissolve into capture errors. In the empirical literature (e.g., Reason, 1984; Schwartz et al., 1998) object substitution errors are differentiated from capture errors through subsequent actions relating to the initial task, not some capturing task. Thus, patient HH (introduced by Schwartz, Reed, Montgomery,

Palmer, & Mayer, 1991) made numerous substitution errors, including preparing a cup of coffee by using a cereal bowl (instead of the cup provided) for the entire task (M. F. Schwartz, personal communication, May 3, 2006). It continues to remain unclear how this behavior could arise within the SRN model.

Anticipation Errors

Botvinick and Plaut (2006) argued that the specific anticipation error of pouring from a sealed container does not arise in the original SRN model's output because the model was never trained on a sequence featuring a container that did not need to be opened. This argument only serves to highlight the importance of the training set in shaping the model's errors and further undermines the use Botvinick and Plaut (2004) made of results from empirical studies in evaluating the model. For example, there can be no guarantee that the effect of lesion severity on error type as reported in, say, Figure 15 of their 2004 article would hold if the model was trained on all sequences to which a typical human is exposed.

Frequency Matching

Botvinick and Plaut (2006) accepted our suggestion of incorporating the Luce choice rule (Luce, 1963) into the procedure for selecting actions on the basis of the SRN model's output. This would address our specific concern about how the model might, when preparing coffee, first add sugar from the packet on 25% of attempts, sugar from the bowl on 25% of attempts, and cream on 50% of attempts, without employing a nondeterministic *fixate-sugar* action. This situation does, however, highlight a fundamental limitation of the SRN model: The selection between these alternatives would be random. It cannot be influenced, for example, by a specific intention to use the sugar bowl rather than the sugar packet. It is also unclear whether frequency matching will scale. Instead, it is conceivable that with more overlapping tasks, use of the Luce choice rule would result in high rates of switching between similar tasks.

Interchangeable Subsequences

In an effort to demonstrate that the SRN is capable of spontaneously generalizing a subsequence from the context in which it was learned to another, on the grounds that another subsequence has been seen to apply in both contexts, Botvinick and Plaut (2006) presented details of an additional simulation with a reduced version of the SRN model. In our view, the demonstration is unconvincing. The behavior of the simulation demonstrating transfer can be explained if two regularities in the training set are learned: that the first and fourth actions are always *use-X* and *fixate-X* respectively, where X is a_1 to a_5, and that the second and third actions are always *fixate-Y* and *use-Y* respectively, where Y is the object in peripheral vision. These regularities hold of every sequence on which the model was trained, the relevant subsequences are both just two actions long, and only one action is required following the subsequence. The domain has therefore been simplified to the extreme. Furthermore, the model needs five versions of the a_i to be successful; it cannot manage with only two. It is therefore an exaggeration to suggest that this simulation "demonstrates, in a simple way, that the SRN model can infer that 2 subsequences may

be used interchangeably in a context in which it has not been directly trained to do so" (Botvinick & Plaut, 2006, p. 919), if the interchangeability is to occur in other than a restricted set of situations.

Variations in Initial Conditions

The SRN model can, Botvinick and Plaut (2006) argued, respond appropriately to variation in the initial state of its environment (e.g., whether the sugar bowl is open or closed) if similar variation occurs in its training corpus. By way of illustration they claim that the SRN as originally trained does adjust its behavior appropriately if the cup initially contains cream. We attribute this success to the facts that (a) following the addition of sugar in such a case, the fixated and held objects precisely match those expected by the model prior to drinking (i.e., at Step 32 in the standard coffee task) but not prior to adding cream, and (b) on 75% of training trials adding sugar was followed immediately by drinking. In fact, what appears to be happening here is that the SRN is actually committing a capture error. Indeed, the model does not stir the cream as it should.

To consider this point further, we also tested the original SRN model with the cup initially containing coffee. The model invariably added more coffee, but on some rare occasions it omitted sugar or both sugar and cream, simply proceeding directly from adding coffee to drinking. The example provided by Botvinick and Plaut (2006) of handling variation in initial conditions is the only variation with which the SRN, as originally trained, can deal adequately.

Catastrophic Interference

The favored solution of Botvinick and Plaut to catastrophic interference is to use an analogy with McClelland, McNaughton, and O'Reilly's (1995) approach to semantic memory. However, they only countered our concern that this requires the hippocampus (required to implement McClelland et al.'s approach) to reliably order extremely long sequences (up to 37 elements in the coffee case) by changing their position and suggesting that pairs of single steps may be adequate (following Ans, Rousset, French, & Musca, 2004). However, many details need to be clarified. Thus, how the sequence would be correctly reconstructed is not transparent and the Ans et al.-type procedure would not be perfect for 37 pairs. Moreover, we reiterate that the anatomical links between the habit system and the hippocampus are more tenuous than between temporal structures associated with the semantic system and the hippocampus, and which support McClelland et al.'s account in the semantic domain. In our view, chunking sequences into their constituent subsequences (made possible by the use of hierarchical structure) offers a more realistic alternative. Learning can then effectively exploit existing knowledge of constituent subtasks.

Wider Issues

The Role of the Training Set Revisited

An unarticulated premise in the argument of Botvinick and Plaut, and a key point made by Cooper and Shallice (2006), is that for the network to generate relevant behavior it must be trained on appropriate sequences—for instance, sequences that legitimize

omission and reversal through their local transition probabilities. Thus, if in the training set *c* always follows *b*, which always follows *a* within some context, the probability of the omission of *b* in that context will be greatly enhanced if some similar context licenses the direct transition of *a* to *c*. Moreover, they did not address the problem that in their approach all errors are essentially blends in which behavior under the guidance of one schema drifts or is captured by another, leaving the system completing a different schema from the one started.

The Role of Goals Revisited

Goals within the interactive activation network (IAN) model are more important than Botvinick and Plaut (2006) suggested. First, if two schemas achieve the same goal and that goal is a subgoal of the current schema, then all else being equal, either may be performed. Thus, the use of goals captures the fact that some schemas have a common purpose. In the case in which the effect of interchangeability is achieved by the SRN model, it is through a substitute for a goal, namely a nondeterministic perceptual action, that of *fixate-sugar*, which results in fixation passing to either the sugar packet or the sugar bowl. Even here though, one finds a lack of generalization as shown in our original Simulation 3. In addition, some errors that occur in neurological patients strongly suggest that low-level action is goal directed. Thus, "body-part-as-tool" errors are not infrequent in the everyday action of ideational apraxic patients (cf. Rumiati, Zanini, Vorano, & Shallice, 2001). These errors involve use of a body part (e.g., a finger) instead of an available tool (e.g., a butter knife) in the performance of a task or subtask (e.g., spreading jam on bread). Critically, action in these cases remains goal directed and the goal is achieved (i.e., jam is spread on the bread), though it is achieved in an atypical way.

Routine and Nonroutine Action

Contrary to Botvinick and Plaut's (2006) claim, we did not make the "assertion" that representations in both routine and nonroutine systems are isomorphic, and we certainly do not believe that they are. Even within a contention scheduling system, the degree of contextual sensitivity that representations need to incorporate may differ according to their level in the schema hierarchy. At the same time, there must be some informational relation between the systems so that the higher system can (a) appropriately instruct the lower system in case of initiating a habit, (b) modulate the lower system in case of recovering from error, and (c) modulate the system so as to perform only a part of, or a nonroutine modification of, a habit (e.g., making coffee with more sugar than usual). In its present guise, the SRN model is able to take two instructions: to prepare and drink coffee or to prepare and drink tea. Our argument is that if the routine system can also, for example, prepare a beverage for someone else (without drinking it), prepare tea with lemon, add sugar to an existing beverage, or to perform any subtask of beverage preparation, then additional instruction units will be required. It then quickly becomes clear that these instruction units function in a way that is isomorphic to goals within the interactive activation network model.

Conclusion

Botvinick and Plaut (2006) have not addressed our fundamental concerns with their eliminativist stance. In particular, in at least three of the areas where they challenge our criticisms (*Object Substitution Errors, Interchangeable Subsequences, Variations in Initial Conditions*), there still appear to be major problems for their model. In addition, they have not demonstrated that the modifications they suggest can in fact be implemented simultaneously in a single computational account that addresses each of our concerns while still being able to acquire tasks, interface with higher cognitive systems, and scale to the range of routines available to adult humans without catastrophic interference during training. Our criticisms, however, relate to their use of unstructured representations. To reiterate, it is our view that these criticisms are not necessarily intrinsic to the SRN approach per se, and it remains to be seen whether they can be addressed through the additional use of hierarchically structured goal-directed schema representations.

References

Ans, B., Rousset, S., French, R. M., & Musca, S. (2004). Self-refreshing memory in artificial neural networks: Learning temporal sequences without catastrophic forgetting. *Connection Science, 16,* 71–99.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review, 111,* 395–429.

Botvinick, M., & Plaut, D. C. (2006). Such stuff as habits are made on: A reply to Cooper and Shallice (2006). *Psychological Review, 113,* 917–928.

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review, 113,* 887–916.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–187). New York: Wiley.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and the neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

Reason, J. T. (1984). Lapses of attention in everyday life. In W. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 515–549). Orlando, FL: Academic Press.

Rumiati, R. I., Zanini, S., Vorano, L., & Shallice, T. (2001). A form of ideational apraxia as a selective deficit of contention scheduling. *Cognitive Neuropsychology, 18,* 617–642.

Schwartz, M. F., Montgomery, M. W., Buxbaum, L. J., Lee, S. S., Carew, T. G., Coslett, H. B., et al. (1998). Naturalistic action impairment in Closed Head Injury. *Neuropsychology, 12,* 13–28.

Schwartz, M. F., Reed, E. S., Montgomery, M. W., Palmer, C., & Mayer, N. H. (1991). The quantitative description of action disorganisation after brain damage: A case study. *Cognitive Neuropsychology, 8,* 381–414.