Psychological Review
2006, Vol. 113, No. 4, 917–928

0033-295X/06/$12.00 DOI: 10.1037/0033-295X.113.4.917

COMMENTS

Such Stuff as Habits Are Made On: A Reply to Cooper and Shallice (2006)

Matthew M. Botvinick
University of Pennsylvania

David C. Plaut
Carnegie Mellon University

The representations and mechanisms guiding everyday routine sequential action remain incompletely understood. In recent work, the authors proposed a computational model of routine sequential behavior that took the form of a recurrent neural network (M. Botvinick & D. C. Plaut, 2004). Subsequently, R. P. Cooper and T. Shallice (2006) put forth a detailed critique of that work, contrasting it with their own account, which assumes a strict hierarchical processing system (R. P. Cooper & T. Shallice, 2000). The authors respond here to the main points of R. P. Cooper and T. Shallice's (2006) critique. Although careful and constructive, the arguments offered by R. P. Cooper and T. Shallice (2006) mistook several superficial implementational issues for fundamental theoretical ones, underestimated the computational power of recurrent networks as a class, and in some ways mischaracterized the relationship between the accounts they compare. In responding to these points, the authors articulate several key theoretical choices facing models of routine sequential behavior.

Keywords: sequential action, computational modeling, procedural memory

The knowledge structures, or schemas, that guide everyday action routines have been of fundamental interest to psychologists since the time of William James (James, 1890; Miller, Galanter, & Pribram, 1960; Norman, 1981; Reason, 1990; Schank & Abelson, 1977). However, whereas the topic has become increasingly central within artificial intelligence (e.g., Barto, Singh, & Chentanez, 2005), human factors research (e.g., John, 2003), and neuroscience (e.g., Grafman, 1995), within psychology it appears to have drifted toward the sidelines. Although important work has certainly continued, both in psychology and in neuropsychology (e.g., Altmann & Trafton, 2002; Buxbaum, Schwartz, & Montgomery, 1998; Forde & Humphreys, 2002; Zacks & Tversky, 2001), research into the representations underlying routine sequential behavior stands in need of reinvigoration. Hence, the series of articles by Cooper and Shallice (Cooper & Shallice, 2000; Cooper, 2003, in press; Cooper, Schwartz, Yule, & Shallice, 2005) introducing an explicit symbolic computational model of routine sequential action is of considerable importance. Inspired by this work, we recently offered an alternative model based on a recurrent neural network architecture (Botvinick & Plaut, 2002, 2004). Cooper and Shallice (2006) have analyzed the relationship between the respective accounts, arguing broadly in favor of their approach and against ours.

Although the comments offered by Cooper and Shallice (2006) rightfully restored the spotlight to some fundamental questions concerning routine sequential action, they also have called for a response. The investigators claimed to demonstrate several specific deficiencies in the behavior of the Botvinick and Plaut (2004) model, but many of these can be shown to stem from fairly superficial implementational factors rather than from basic theoretical assumptions. In translating their specific observations into general statements about the Botvinick and Plaut (2004) framework, Cooper and Shallice underestimated that framework's computational capacity and overestimated its dependence on ad hoc assumptions. Finally, in characterizing the relations between the Botvinick and Plaut (2004) account and their own, Cooper and Shallice drew some debatable distinctions while at the same time downplaying some critical ones.

In the present article, we reply to Cooper and Shallice (2006). Although we aim to rebut a set of specific claims, our broader goal is to advance the debate by laying out some key theoretical issues that are raised by the contrast between the relevant models but which have not yet been adequately articulated. These bear primarily on two questions: (a) What kind of representational medium supports routine sequential behaviors? and (b) What is the computational basis of generativity in routine sequential behavior?

Our reply is organized into four sections. In the first, we consider the specific results that Cooper and Shallice reported from simulations using the Botvinick and Plaut (2004) model. In the next section, we address two broad criticisms Cooper and Shallice made of that model, both of which relate to the issue of generativity. In the third section, we offer an alternative perspective on what Cooper and Shallice identified as the key differences between

Matthew M. Botvinick, Center for Cognitive Neuroscience, University of Pennsylvania; David C. Plaut, Department of Psychology, Carnegie Mellon University.

The present study was supported by National Institute of Health Awards MH16804 to Matthew M. Botvinick and MH64445 to J. McClelland and David C. Plaut. Code for the simulation presented in the Appendix is available for download at www.ccn.upenn.edu/~mmb

Correspondence concerning this article should be addressed to Matthew M. Botvinick, Center for Cognitive Neuroscience, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 10104-6241. E-mail: mmb@mail.med.upenn.edu

their paradigm and the one proposed in Botvinick and Plaut (2004). Finally, we consider the parallel Cooper and Shallice drew between the present debate and the rules versus connections debate currently underway in some other domains of inquiry.

Throughout the article, in keeping with Cooper and Shallice (2006), we refer to their computational framework as the *IAN* (Interactive Activation Network) *model* and to the Botvinick and Plaut (2004) model as the *SRN* (Simple Recurrent Network) *model*.

Specific Observations

Cooper and Shallice (2006) presented observations from a series of simulations using the SRN model that are purported to indicate flaws in the theory the model implements. Six observations are considered to be of particular concern: (a) the infrequency of object substitution errors; (b) the failure of the model to produce one specific anticipation error; (c) the failure of the model to match, at test, the frequencies of tasks during training; (c) the failure of the model to infer that certain subsequences are equivalent and interchangeable; (d) the failure of the model to cope with novel initial conditions; and (e) the susceptibility of the model to catastrophic interference. In the following sections, we revisit these issues, arguing that the observations reported in Cooper and Shallice (2006) can largely be attributed to relatively inconsequential implementational considerations rather than to core paradigmatic assumptions.

Object Substitution Errors

Object substitutions, as observed both in normal behavior and in apraxia, involve incorrect use of one object in place of another—Cooper and Shallice (2006) gave the example of pouring instant coffee grounds into a sugar bowl instead of a cup. As Cooper and Shallice (2006) noted, such errors do occur in the SRN model (see Botvinick & Plaut, 2004, p. 417). However, as demonstrated in their Simulation 1, substitutions make up a smaller proportion of all errors than observed in human behavior.

As Cooper and Shallice (2006) correctly noted, for the SRN model to acquire the wrong object in a given context, the model must explicitly select that object, and it is unlikely to do this unless its internal representation of task context is already disrupted. Thus, having acquired the wrong object, the model is more likely to select an action appropriate to that object than it is to use the object to execute the correct action for the task context.

However, the infrequency of object substitution errors in the SRN model stems from an inessential implementational choice. For simplicity, Botvinick and Plaut (2004) modeled object selection as a deterministic process; when the model selected the *fixate-cup* action, the cup reliably appeared on the next time step as the viewed object. This choice was made to minimize the potential sources of error in the model so as to emphasize the role of the model's internal context representations in generating errors. However, it is obvious from everyday life and well documented in empirical studies (Zhang, Samaras, Yang, & Zelinsky, 2005) that object selection is not deterministic. Searching for one object often leads to the inadvertent and typically transient selection of another. Incorporating this fact into the SRN implementation would have increased the frequency of object substitution errors.

This can be illustrated by examining the behavior of the original Botvinick and Plaut (2004) model if the incorrect visual input is presented following a search action. We ran the model on the coffee-making task used in Botvinick and Plaut (2004) until it had torn open the coffee packet and selected the output *fixate-cup* in preparation to pour.[1] However, rather than providing the cup as the viewed object on the next time step, we instead presented the sugar bowl (input features *cup-shaped*, *two-handles*, *sugar*). Faced with this input, on most trials (78%), the model selected *pour* (*fixate-spoon* was selected on 17% of trials and *fixate-cup* on the remaining 5%). On the basis of the classification introduced by Schwartz, Reed, Montgomery, Palmer, and Mayer (1991), this error—pouring the coffee grounds into the sugar bowl—would count as a substitution error.

Of course, if the model were trained under circumstances in which search occasionally yielded the incorrect object, it would presumably be less prone to error under such circumstances. Nevertheless, under noise, such a change would obviously lead substitution errors to become more frequent than in the original model.[2]

Anticipation Errors

A second type of action error that Cooper and Shallice (2006) focused on is the anticipation error, in which a critical action is skipped. As is acknowledged in Cooper and Shallice (2006), the SRN model does produce such errors at rates resembling those observed empirically. However, it is pointed out that the model fails to commit a particular anticipation error that does occur in the coffee task when performed by patients with action disorganization syndrome. Here, in adding cream, the step of opening the cream container is omitted, leading to an effort to pour from a sealed container.

However, the failure of the original model to emit this error reflects an inessential implementational choice rather than a deep-seated flaw in the paradigm. The error fails to occur for the simple

[1] The Botvinick and Plaut (2004) model was retrained on the coffee and tea sequences as described in the original article. However, the background set was not included. The background set included only scenarios involving the objects involved in the coffee and tasks and so contained no instances of pouring into the sugar bowl, as one might do in refilling. Thus, the background set inappropriately biased against performing this action as an error. Noise was added to hidden unit activations as described in Botvinick and Plaut (2004) at a level of 0.2. All simulation results reported in the body of the article were replicated in five separate training runs.

[2] This, in turn, would also address two other concerns raised by Cooper and Shallice (2006), allowing the presence of distractor objects to affect performance and driving down the proportion of omission errors. In anticipation of further debate, it is noted that the use of "perceptual actions" as outputs of the SRN model is another implementational convenience, standing in for a mechanism whereby objects are selected by top-down biasing on the processing of bottom-up perceptual inputs (see Botvinick, Bylsma, Buxbaum, & Jax, in press).

Note that our argument in the present section, as in subsequent sections, implies the assumption that introducing the specified changes to the Botvinick and Plaut (2004) model would leave previously established aspects of the model's behavior intact. Given that we have not tested this through detailed simulations, the assumption can be questioned. Nevertheless, in the absence of obvious causes for doubt, the assumption appears justifiable.

reason that, in the repertoire of tasks on which the model was trained, there is no example of a container that does not require opening before use. Thus, the sequence *pickup→fixate-cup→pour* never occurs in any context during training. In the real world, containers are often open when first encountered (think of adding cream to coffee from a pitcher), making the sequence *pickup→fixate-cup→pour* quite familiar and more likely to intrude as an error. Indeed, if the model is trained on coffee sequences in which the cream is sometimes already open when first encountered and then tested at a noise level of 0.2, it does occasionally commit the anticipation error that is of interest to Cooper and Shallice (2006).

Frequency Matching

Simulation 2 of Cooper and Shallice (2006) demonstrated that the SRN model, as implemented in Botvinick and Plaut (2004), does not produce sequences at the same relative frequencies with which those sequences occurred during training. For example, the simulation showed that if the model is trained with a slight frequency bias toward adding cream before sugar, the trained model shows a much stronger bias. Cooper and Shallice (2006) concluded that "the frequency of sequences in the training set must be finely balanced if the SRN model is to be able to generate all sequences on which it has been trained" (p. 895).

This aspect of the model's performance stems from another inessential implementational decision, which was to use a winner-take-all method for output selection. As Cooper and Shallice (2006) pointed out (p. 902), an equally reasonable choice would have been to select actions probabilistically, with the likelihood of an action depending on the activation of the relevant output unit (above some minimum). In fact, such a selection rule was eschewed by Botvinick and Plaut (2004) simply for expository reasons. We wished to avoid introducing another potential source of errors, again reflecting a desire to make clear the effect of degrading the model's context representations. Had a probabilistic selection procedure been used, such as the Luce choice rule (Luce, 1963), it is evident that the model would come closer to matching the frequencies of sequences in the training set, a point that Cooper and Shallice (2006) themselves allowed (p. 902).

Interchangeable Subsequences

An important feature of routine sequential action is that it often contains subsequences that can be substituted for one another. Botvinick and Plaut (2004) showed that the SRN model can learn to treat subsequences interchangeably, learning specifically to add sugar from a sugar packet or a sugar bowl. However, Cooper and Shallice (2006) raised the question of whether the model can infer such sequence equivalence. Specifically, Simulation 3 demonstrated that if the model is trained to prepare tea by using both a sugar packet and sugar bowl but to prepare coffee by using only a sugar bowl, the fully trained model does not spontaneously produce coffee sequences in which the sugar packet is used.

Although it is true that Botvinick and Plaut (2004) did not directly demonstrate the ability to infer sequence equivalence, that article did note that SRN models are capable of it and referred to unpublished simulation results documenting this point (p. 423–424). The relevant simulation, which we now briefly describe, and which is further detailed in the Appendix, also makes clear why Cooper and Shallice (2006) failed to observe the desired effect.

To keep things simple, our simulation posits a formalized task domain that boils the issue down to its basics. It is assumed that all legal action sequences within this domain have the same structure. Each starts with the use of some object *a*, transitions to the use of a second object *b*, and then finishes with a return to the original object *a* (see the Appendix for details of the task and model implementation). A critical assumption is that there are two objects, b_1 and b_2, analogous to the sugar packet and sugar bowl, that can fill the *b* role. It is assumed, additionally, that b_1 and b_2 can be used interchangeably in a variety of different contexts, and to capture this we assume that five different objects ($a_{1\text{-}5}$) can fill the *a* role. As a result, there are 10 legal sequences in the domain: $a_1 \rightarrow b_1 \rightarrow a_1$, $a_1 \rightarrow b_2 \rightarrow a_1$, $a_2 \rightarrow b_1 \rightarrow a_2$, $a_2 \rightarrow b_2 \rightarrow a_2$, and so forth. In our simulation, the model is trained on all but 1 ($a_1 \rightarrow b_2 \rightarrow a_1$). The question of interest is whether, following training, the model produces that withheld sequence. As detailed in the Appendix, when trained and tested under a reasonable set of assumptions, the model does produce this sequence. This 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.

A critical observation from the above simulation is that the model fails to generalize in the same way if a more restricted training set is used. In the training set just described, both versions of the object-*b* subsequence occurred in four different contexts ($a_{2\text{-}5}$). However, if sequences involving $a_{3\text{-}5}$ are removed from the training set, so that the model observes both b_1 and b_2 in only one context (a_2), the trained model's behavior does not reflect the inference that the two subsequences can be used interchangeably in the a_1 context. Details of the relevant simulation are once again provided in the Appendix. The crucial point is that, to generalize in the fashion Cooper and Shallice (2006) quite reasonably demanded, the SRN model must be exposed during training to an adequate range of sequence variation. Restricting the training set, as must be done for practical reasons, necessarily exposes the network during learning to spurious correlations that are unlikely to arise in actual human experience—such as the invariable and massively repeated use of one of two permissible methods for adding sugar during tea making, as in the Cooper and Shallice (2006) simulation (for related computational observations, see Rougier, Noell, Braver, Cohen, & O'Reilly, 2005). The reliance of the SRN model on a training set that is broad and representative of the task domain is a point we shall further emphasize in what follows.

Variations in Initial Conditions

Another claim from Cooper and Shallice (2006) is that the SRN model performs reasonably only if presented with precisely the same environmental conditions as encountered during training. Given different circumstances, it is suggested, the model cannot infer correct modifications to action sequences. The example provided in Simulation 4 of Cooper and Shallice (2006) involved running the trained model on the coffee task but also initializing the environment such that the sugar bowl is initially open, a situation never encountered during training. The sensible thing to do when encountering a sugar bowl with a spoon in hand (as the

model does) is of course to go ahead and scoop. Instead, the model put down the spoon and momentarily picked up the sugar bowl before picking up the spoon again, scooping sugar and reentering the standard coffee sequence. As bizarre as the model's behavior may appear in this case, its implications for the underlying theory are not as significant as Cooper and Shallice (2006) suggested.

What Cooper and Shallice (2006) were asking the model to do in this simulation is to exhibit a form of generalization. As the previous section concluded, before such generalization can occur, the model must be exposed to an adequate, though not exhaustive, range of variation. The model's behavior in the Cooper and Shallice (2006) simulation is a direct consequence of the use of a highly restricted training set. As mentioned earlier, the Botvinick and Plaut (2004) model never encountered during training any container that did not need to be opened before use. There is every reason to expect that if the model were trained on a wider range of task sequences, including tasks involving the use of uncovered containers, that it would show precisely the kind of generalization Cooper and Shallice (2006) sought.

The point can be substantiated by pointing to other equally meaningful instances in which the model does behave sensibly in the face of novel initial conditions. For example, if the environment is initialized so that the coffee cup already contains cream at the beginning of the trial—a situation never presented during training—the model adds coffee grounds and sugar but skips the cream-adding sequence. In this instance, the contents of the training set provided an adequate range of sequence variation for the model to infer that cream need not be added if it is already seen to be in the cup.

Catastrophic Interference (CI)

A well-known property of neural network models that use distributed representations is their susceptibility to CI, the disruption of initial learning by later training on a different task (French, 1999). Cooper and Shallice (2006) demonstrated, in Simulation 5, that the SRN model is no exception: To learn two tasks, the model must be trained on both in an interleaved fashion. The susceptibility of the SRN model to CI was acknowledged in Botvinick and Plaut (2004), but it is not the fatal flaw that Cooper and Shallice (2006) suggested. In particular, McClelland, McNaughton and O'Reilly (1995) have suggested how CI might be avoided through the interaction of dual learning systems, one based on medial temporal lobe structures including the hippocampus. Cooper and Shallice rejected the dual system theory as "neuroscientifically and cognitively implausible" (p. 906), at least as an account of how action sequences are learned. In particular, they state that there "is no evidence that the hippocampus can retain and order completely accurately a very long sequence of input-to-output mappings" (p. 896). However, there is abundant evidence that the hippocampus is involved in encoding sequences (see Botvinick & Plaut, 2004, p. 424), and there is no evidence of which we are aware to suggest that there is a hard limit on the capacity of hippocampal sequence memory, falling below the relevant sequence length. Furthermore, it might not be necessary for medial temporal lobe memory systems to encode long sequences; as Ans, Rousset, French, and Musca (2004) have demonstrated, CI in sequence learning can be prevented by mechanisms that generate only single-step input–output mappings.[3]

Cooper and Shallice (2006) further attempted to undermine the idea that the hippocampal system may alleviate CI by commenting on the connectivity in rats between hippocampus and subportions of the dorsal striatum thought to subserve habit production. However, the medial temporal cortex is well known to interact widely with neocortex, including areas throughout frontal, parietal, and temporal cortex that are certain to be involved in representing the perceptual inputs and associated actions involved in habits. Furthermore, the hippocampal complex also interacts with regions of the prefrontal cortex generally agreed to support planning and action in nonroutine situations (Cohen & O'Reilly, 1996). Bringing in the McClelland et al. (1995) dual learning systems theory thus in no way contradicts the proposal in Cooper and Shallice (2006) that the learning of action routines is mediated in part by a higher-level action system dedicated to the programming of actions in nonroutine circumstances. Nevertheless, unlike Cooper and Shallice (2006), we do not assume that habitual action sequences are always or primarily learned via such a route. In a reductio ad absurdum of the Botvinick and Plaut (2004) theory, Cooper and Shallice (2006) argued that hierarchically structured action sequences are not learned through "unguided imitation or observation of lengthy, apparently purposeless action sequences" (p. 899). However, there is abundant empirical evidence that people actually are quite good at abstracting the sequential structure of purposeless sequences (Avrahami & Kareev, 1994; Cleeremans, 1993; Saffran, 2001), and this capacity is typically attributed to systems underlying procedural memory, systems that seem likely to support routine sequential action in everyday life (Poldrack, Prabakharan, Seger, & Gabrieli, 1999).

Broader Criticisms of the SRN Model

The preceding sections have shown that many of the problems Cooper and Shallice (2006) purported to reveal in the SRN model can be addressed without substantially altering the underlying account put forth by Botvinick and Plaut (2004)—in most cases by simply implementing the account in greater detail. Having responded to specific observations, we turn now to two key generalizations that Cooper and Shallice (2006) advanced, on the basis of those observations, concerning the theme of *generativity*. The first is the claim that the SRN model cannot produce sequences it has not been directly trained upon; the second is the related claim that the behavior of the model is unreasonably dependent on the structure of the training set.

Generativity

Cooper and Shallice (2006) claimed that the SRN model "can only produce—even as errors—sequences of actions on which it has been substantively trained" (p. 905). With regard to errors, this is clearly untrue. As indicated in the section *Anticipation Errors*, the model does produce errors involving action sequences that do not occur in the training set. Indeed, Botvinick and Plaut (2004) reported such errors (e.g., pouring directly from the sugar bowl into the cup, p. 417). It is true that the errors produced by the SRN

[3] Importantly, this work by Ans et al. (2004) did address non-Markov sequences, that is, sequences whose reproduction requires preservation of temporal context information.

model are influenced by the structure of the training set, in the sense that the model is most prone to errors that fit with any regularities of sequential structure that characterize the training set. However, far from being a flaw of the account, this is in fact one of its strengths, because the same characteristic has been observed empirically in studies of human slips of action (Reason, 1990), as in work on speech errors (Dell, Reed, Adams, & Meyer, 2000).

Neither is the Cooper and Shallice (2006) claim true with regard to correct sequences. As shown above, in the sections *Interchangeable Subsequences* and *Variations in Initial Conditions*, the SRN model is quite capable of producing legal sequences that were never presented during training. The generative capacity of recurrent neural networks is further demonstrated by previous research in which such models have been used to compose music. Networks trained on a variety of melodies have been shown in several studies to generate novel tunes in the same musical style, melodies not matching any of those on which the model was trained (Eck & Schmidhuber, 2002; Mozer, 1994). This work clearly demonstrates that SRNs can generate sequences that are consistent with the structure implicit in a training set, without matching any exemplar presented during training. A critical point is that this sort of generalization requires the training set to contain a broad and representative sample over the relevant domain. The music-composing SRN need not be exposed to all possible melodies in a given style, but it does require exposure to a legitimate number of them. The importance of a representative—though not necessarily exhaustive—training sample is the same one we emphasized above in discussing interchangeable sequences.

Importance of the Training Set

Cooper and Shallice (2006) argued repeatedly that the behavior of the SRN model depends unreasonably on the detailed structure of the chosen training set:

> the training set has to be fine-tuned to produce the appropriate output . . . [thus] the negative property of being hand coded is merely transferred from the weights for the IAN model to the training set for the SRN model. (pp. 905–906)

Again, it is true that the behavior of the SRN model is shaped by the structure of the sequences presented during training, and this was highlighted by Botvinick and Plaut (2004) as one of the core tenets of the proposed account. Regrettably, Botvinick and Plaut (2004) made no direct comment on the specific form of the training sets used in the simulations it reported. However, two key assumptions were strongly implied. The first assumption is that the sequences observed by the system during learning constitute a representative sample of well-formed behavior in the relevant tasks. This means, above all, that what can vary in an action routine will vary in the sequences observed during learning, although—as just explained—the training set need not contain all well-formed sequences but merely a sufficient sampling. The second assumption (implied on p. 403 of Botvinick & Plaut, 2004) is that learning of any specific task takes place within the context of learning a very broad variety of other tasks and that there are regularities of sequential structure spanning multiple tasks that support generalization across task lines.

For practical reasons, it was not possible to fully implement the second of these assumptions in the simulations reported in Botvinick and Plaut (2004). Indeed, those simulations are quite artificial in that they attempt to model the tasks of coffee and tea making without directly modeling the thousands of other tasks a typical Westerner would also have in his or her repertoire. This accounts for why, as Cooper and Shallice (2006) found, the model's behavior is so sensitive to the precise form of the coffee and tea training sets, and why the model does not generalize in all of the ways one would expect of a human actor. Because the training set was necessarily small, its details had an overly large influence on the behavior of the model, in precisely the same way that individual observations in a small sample can unduly influence the fit of a statistical model. However, as with statistical models, the robustness of learning and the capacity for generalization increase with the sample size. A strong assumption of the Botvinick and Plaut (2004) account, which we consider to be face valid, is that human learners are exposed to a very broad and representative range of action sequences.

Paradigmatic Assumptions: Schemas, Hierarchy, Goals

One strength of Cooper and Shallice (2006) is that it goes beyond a simple critique of the SRN model, attempting to identify the fundamental theoretical differences between the IAN and SRN models. According to Cooper and Shallice (2006), the main points of contrast center on the purportedly "eliminativist" stance of the SRN model with respect to three elements said to be defining of the IAN framework: (a) explicit schema representations, (b) hierarchical network structure, and (c) explicit goal representations. Although this framing pinpoints a critical set of issues, we would propose a somewhat different perspective on the role of schemas, hierarchies, and goals in routine sequential action.

Schemas

A core assumption of the IAN model is its inclusion of *schema nodes*, elementary components or units that are identified with entire tasks or subtasks. Cooper and Shallice (2006) argued that it is, in fact, necessary to assume such atomic computational elements to account for specific aspects of human behavior. These include anticipation errors, adjustment to novel environmental conditions, and generation of well-formed but novel sequences (see Cooper and Shallice, 2006, pp. 893–894). However, as demonstrated earlier, the SRN model can address all of these without assuming localist schema representations. The SRN model develops, through learning, the functional equivalent of the IAN model's schema nodes. Rather than being represented by single units, task and subtask identity is represented in a distributed fashion, forming part of the internal context representation analyzed in Botvinick and Plaut (2004).

Whereas it is too strong to say, as Cooper and Shallice (2006) did, that the SRN model is eliminativist with respect to task and subtask representations (i.e., schemas), it is true that the relevant patterns of activation may be more difficult to isolate within the SRN model than in the IAN model. Intuition may suggest that this is a problem. After all, in daily communication we use terms that refer to whole sequences of behavior, abstracting over their details: going to the movies, playing a game of tennis, checking email, and so forth. In the artificial intelligence literature, such high-level representations are referred to as *temporal abstractions* (Barto &

Mahadevan, 2003; Barto et al., 2005; Sutton, Precup, & Singh, 1999), and there has been growing interest in their use. We consider it informative, however, that the primary importance of temporal abstractions within artificial intelligence has to do with their role in reducing the search problem involved in planning, rather than any role in supporting the execution of established routines. This seems to reflect a point that is equally applicable to human behavior: In planning one's day, it may be useful to have a compact representation of going to the store that abstracts over and is quite independent of the internal details of this routine. However, in actually executing the routine, such abstraction is not particularly important, and for reasons that we shall discuss in the next section, it may in fact be counterproductive. On the basis of these considerations, we consider it likely that the generation of human behavior does leverage highly abstract task representations but that these are put to use by the processes involved in the planning and programming of nonroutine action sequences rather than by the processes underlying habit production, which use context representations tied in a more intimate way to real-time execution.

Hierarchical Structure

A key assumption of the IAN model, which is underscored by Cooper and Shallice (2006), is that the processing system underlying routine sequential action is strictly and constitutively hierarchical in structure. Cooper and Shallice (2006) claimed that this assumption is necessary on the basis of arguments similar to those used to defend schema nodes. One claim that received particular emphasis (p. 899) relates to the ability to infer subsequence interchangeability, an ability that Cooper and Shallice (2006) believed the SRN model to lack, as explained earlier. Cooper and Shallice (2006) suggested that the ability to transfer a familiar subtask sequence into a task context in which it has never previously appeared requires that the subtask be represented in a way that is invariant with respect to context. Thus, the claim goes, the routine for adding sugar must be represented in precisely the same way, regardless of whether it appears in the context of coffee or tea making.

To some degree, this point must be valid. Clearly, subtask representations must be separable from higher-level task representations if subtasks are to be performed in multiple contexts. There is no barrier to this in the SRN model; as demonstrated earlier, with sufficient variability in the training set, the model does develop the ability to transfer subsequences to new contexts. However, it is worth considering whether absolute context independence, as required in the IAN framework, is a desirable representational property. It is a conspicuous feature of naturalistic human behavior that the way a subroutine is performed often depends on the larger task context in which it occurs (Agre, 1988). For example, the amount of sugar one adds to a beverage may depend on whether the beverage is coffee or tea. As we have observed in previous work (Botvinick & Plaut, 2002, 2004), this context dependence raises a difficulty for strict hierarchical computational accounts: Should sugar adding be represented by one schema or two? If two schemas are assumed, this ignores the fact that different versions of sugar adding are likely to share a great deal of structure (see Schank & Abelson, 1977). If one, then how is execution to be modulated by context?

Cooper and Shallice (2006) appeared to acknowledge that this dilemma strains the limits of the IAN account. To deal with it, they suggest that the framework could be extended by including an inheritance mechanism by which parameters of lower-level schemas could be controlled by parent schemas (p. 895). Whereas this might mitigate the problem, it would also take the account one step closer to resembling a full-fledged programming language, and it would amount to another case in which the theory simply stipulates what it is intended to explain. Moreover, there is empirical evidence that action representations at the neural level are far more context dependent than the IAN model assumes. Specifically, Aldridge and Berridge (1998) observed in rats that the set of basal ganglia neurons active during specific grooming movements differed dramatically depending on whether the relevant movement was executed inside or outside the context of the animal's grooming sequence (see also Salinas, 2004).

The issue we are raising here is not merely a technical one, pertaining only to the contrast between the IAN and SRN models. What we are arguing is that, within the domain of routine sequential action, there is an inherent tension between the need for some degree of context independence, responding to the part-whole structure of everyday tasks, and the need for context sensitivity. Any model of routine sequential behavior must provide an account for how a balance is struck between these, accommodating what we have termed the *quasi-hierarchical* structure of everyday action routines (Botvinick & Plaut, 2004). The IAN model, in its current instantiation, attends exclusively to the demand for context independence by assuming a strictly hierarchical and localist representational regime. As a result, the account faces difficulty with the issues of information sharing and context sensitivity, creating the need for additional, ad hoc mechanisms like inheritance.

The SRN model addresses the balance between context independence and context sensitivity at a more basic level. Rather than assuming hierarchy as a strict constraint on representational structure, the framework starts with a large and unstructured representational space that is shaped by experience with specific task repertoires. Where independence between levels of structure is needed for successful control, the model is entirely capable of developing internal representations that capture this independence, as demonstrated by the simulation of subtask equivalence described earlier. However, because the model's representations are not constrained to be strictly hierarchical, there is room for representational overlap when tasks share structure, and for interactions across levels of task structure, whatever form these may take.

This point relates to the question, raised by Cooper and Shallice (2006), of whether it might be possible to reduce the SRN model to the IAN account. Cooper and Shallice (2006) offered the interesting and valid observation that some distributed models can be reduced to localist models, whereas others cannot. However, they mistakenly characterized the SRN model as representing an "extreme nonreductionist position" (p. 891). In our view the SRN model relates to the IAN model as special relativity relates to Newtonian physics. The former reduces to the latter in the limit of small speeds. Analogously, the SRN model reduces to the IAN model in the limit of strictly hierarchical task structure. To unpack this point, consider that a system can be reduced from distributed to localist when the system's representations are mutually orthogonal. In the case of the SRN model, the representations at issue are the activation vectors that indicate task and subtask context. Note

that there is nothing in the system's architecture to prevent these from being orthogonal. Indeed, the tea-making context could in principle be represented by a single hidden unit just as in the IAN model. Thus, at a functional level, the SRN model can implement a hierarchical representational scheme. Indeed, SRNs have been used in some research (see Rodriguez, Wiles, & Elman, 1999) to implement push-down automata, precisely the kind of mechanism used in some production system architectures (e.g., adaptive control of thought—rational [ACT–R]; Anderson & Lebiere, 1998) to support hierarchical goal and subgoal management. However, unlike the IAN model, the SRN model is not restricted to orthogonal representations, allowing it also to accommodate departures from strict hierarchy where they exist in action routines.

Cooper and Shallice (2006) proposed a framework for reconciling the IAN and SRN accounts in which the localist representations in the IAN network correspond to point attractors in a recurrent neural network. We consider this an appealing direction, largely because it relaxes the constraint of strict hierarchy that currently limits the IAN account. Indeed, in recent work, we have reported an implementation of the Botvinick and Plaut (2004) theory that takes the form of an attractor network (Botvinick, in press). This work takes a further step toward reconciling the two accounts by addressing neuroscientific data suggesting that sequential action is supported by neocortical networks that assume a roughly hierarchical organization (Fuster, 1997).

Goals

The third element Cooper and Shallice (2006) identified as distinguishing between the IAN and SRN accounts involves the role of goals. Cooper and Shallice (2006) defined a goal as "a state of affairs that an agent aims to achieve" (p. 888), stating that "a schema may be seen as a means of achieving a goal" (p. 888). Goals are considered to play a critical role in the operation of the IAN model, whereas it is claimed that "goals play no role in the functioning of the SRN model" (p. 898).

However, Cooper and Shallice (2006) overestimated the role of goal representations in the IAN model. At an operational level, the goal nodes in the IAN model are simply gates on activation from high-level schemas to lower levels. As Cooper and Shallice (2006) explained, "When a parent schema is selected, it does not excite all of its component schemas, just those whose preconditions are satisfied and whose post-conditions are not satisfied" (p. 895), and it is goal nodes that enforce the latter constraint. Thus, goals in the IAN framework ultimately function simply as negative preconditions on schemas. Like the "test" component of the classic TOTE unit of Miller et al. (1960), they simply close off certain portions of activation space when particular conditions hold. One can thus exhaustively describe the functional role goals play in the IAN model without any appeal to the notions of *purpose* or *aim*.

Although Cooper and Shallice (2006) claimed that goals play no role in the SRN model, that model's behavior clearly reflects gating by negative preconditions. This is shown, for example, by the fact that the model skips cream adding if cream is already present in the cup (see *Variations in Initial Conditions*) and by the fact that the model continues to execute the *sip* action until the cup is empty (see Botvinick & Plaut, 2004, p. 423). Unlike the IAN model, there is no special structural element dedicated to implementing the relevant precondition gates, but at a functional level the model recognizes "goals" in precisely the same limited way that the IAN does. Cooper and Shallice (2006) suggested that the implementation of goals in the IAN model allows it to cope with certain situations in which the SRN model is said to fail. However, the supposed problems—dealing with interchangeable subsequences and recognizing when environmental conditions make it unnecessary to execute particular actions—we have already been able to set aside.

In the final analysis, there is little if any difference between the IAN and SRN models in terms of the functional role played by goals. Neither model involves goals in the strong sense of that term: In much of psychology and artificial intelligence, the term *goal* typically denotes a representation of a desired outcome that is matched against action effects as part of a process of means–ends analysis. When understood in this sense, goals are tied closely to problem solving or planning, functions that both Cooper and Shallice (2006) and we attribute to systems separable from those supporting highly routine behavior. In our view, the computations underlying routine sequential behaviors do not, in fact, depend on goals in this strong sense of the term.

Indeed, this idea fits precisely with recent work that Cooper and Shallice (2006) cited, delineating the division of labor between *goal-directed* and *habit* systems. On the basis of animal studies, Balleine et al. (Balleine & Dickinson, 1998; Yin, Knowlton, & Balleine, 2004) have concluded that the system underlying routinized behaviors, that is, the habit system, is not driven by representations of desired and anticipated outcomes but instead operates in a purely reactive way (see also Dickinson, 1985). This is also consistent with reinforcement learning accounts of basal ganglia function, which also posit a reactive mechanism for action selection (Daw, Niv, & Dayan, 2005). The SRN model provides an account for how a reactive habit system could give rise to relatively complex action routines, routines that respond flexibly to varying conditions and that reliably bring about specific outcomes, but that do not rely on goals of the kind involved in planning and problem solving.

Relation to Nonroutine Action

Although we concur with Cooper and Shallice on the distinction between two interacting systems underlying action control, a habit (or Contention Scheduling) system and a goal-directed (or Supervisory Attentional) system, one important difference between the Cooper and Shallice (2006) account and our own concerns the assumed relationship between the representations put to use by these two systems. Cooper and Shallice (2006) implied that the representations inhering in the goal-directed and habit systems are essentially identical in nature. This is suggested, for example, by the proposition that the goal-directed system constructs "temporary schemas" (p. 897) that are used as a basis for "instructing the habit system" (p. 899) and that the habit system can be understood as a "plan library" (p. 888). However, it is most directly indicated by the discussion in Cooper and Shallice (2006) of the interaction between goal-directed and habit systems. Consistent with the well-known theory of Norman and Shallice (1986); Cooper and Shallice (2006) assumed that the goal-directed system (identified with the Supervisory Attentional System) operates by providing top-down input to the habit system (identified with the Contention Scheduling system). The Botvinick and Plaut (2004) account is compatible

with this idea, as noted earlier. However, Cooper and Shallice (2006) made the further assertion that, for the goal-directed system to communicate with the habit system effectively, there must be a one-to-one correspondence between schema and goal representations in the two systems. This position is implied in the criticism that the absence of discrete, isolable schema and goal representations in the SRN model "limits the extent to which representations used by the routine system can be communicated to and manipulated by the nonroutine system" (p. 905).[4]

In contrast to this claim, numerous other accounts of action control have suggested that the representations underlying established action routines are different in form from those supporting nonroutine action, or more specifically, planning. This is true, for example, of the account of proceduralization in the ACT–R paradigm (Anderson, 1987), as well as of the neural lesion studies of Balleine et al. (Balleine & Dickinson, 1998; Yin et al., 2004) and related computational work (Daw et al., 2005). Of possible relevance, there is fairly extensive behavioral and neuroscientific evidence for an analogous representational distinction in the domain of spatial navigation (see Hartley, Maguire, Spiers, & Burgess, 2003).

At a quite general level, Rougier et al. (2005) have argued that there is an intrinsic tradeoff in neural computation between the use of graded, distributed representations, which can capture detailed patterns of similarity and overlap, and the use of more abstract and categorical representations, which are more amenable to arbitrary manipulation and recombination. Rougier and colleagues proposed that different brain systems may assume different points on the continuum between these extremes, on the basis of the demands of the particular operations those brain areas perform. This proposal resonates with our own earlier comments concerning the issue of temporal abstraction, in which we suggested that different degrees of abstraction may be appropriate for executing routine procedures and for planning. We consider it an important theoretical difference between the SRN and IAN accounts that the former leaves open the possibility of multiple formats for action representation, whereas the latter commits to a uniform code, spanning habit and goal-directed systems.

Rules Versus Connections in the Habit System

In framing the relationship between their own theory and the one laid out in Botvinick and Plaut (2004), Cooper and Shallice (2006) attempted to draw a parallel to the rules versus connections debate in language, which contrasts symbolic and parallel distributed processing mechanisms (McClelland & Patterson, 2002). It is not clear that this analogy is neatly applicable. After all, the IAN framework is based directly on a connectionist processing architecture (McClelland & Rumelhart, 1981), and, in enumerating the supposedly symbolic components of their theory, Cooper and Shallice (2006) placed links between nodes at the top of the list (p. 889). Moreover, other components described as irreducibly symbolic in nature, specifically pre- and postconditions, we have shown to have direct functional analogues in the SRN model. Finally, the rules versus connections debate, at least as it arises in research on the English past tense, centers in large part on the question of whether there is a single system for the relevant set of transformations or more than one system. In the case of sequential action, it is agreed there seem to be two systems involved, a goal-directed system and a habit system, the latter of which is our focus.

Despite these inconsistencies, the rules versus connections debate does resonate with the present discussion in at least one important way. Like rule-based accounts of past-tense formation (Pinker, 1999), the Cooper and Shallice (2006) account began by focusing on a salient structural characteristic of behavior (i.e., hierarchy) that largely though approximately characterizes the domain and builds this same structure directly into the architecture of the processing system. As with rule-based accounts of past-tense formation, the challenge then becomes to explain how the processing system copes with secondary aspects of behavior that strain or violate the governing structural principle initially assumed. This leads to the stipulation of additional mechanisms to handle such exceptions to the rule. In the case of the Cooper and Shallice theory, the list of additional mechanisms has grown over time and now includes manner and quality features and an inheritance mechanism (Cooper and Shallice, 2006, pp. 896–897), goal decay (p. 904), a type-token distinction that allows multiple instances of a schema to be created (p. 898), and precondition gates that preserve information about previous actions.

The theory laid out in Botvinick and Plaut (2004), which we have further articulated here, takes an approach akin to the one adopted by those pursuing the connectionist side of the rules versus connections debate (Plaut, McClelland, Seidenberg, & Patterson, 1996). The Botvinick and Plaut (2004) account began by assuming an essentially unstructured representational space and a general purpose learning mechanism and then investigated how particular patterns of behavior might emerge from these in the context of a particular environment. As in work on past-tense formation, this approach results in an account that portrays the most systematic aspects of behavior as emerging out of a system that can also accommodate finer-grained aspects of behavior that cut across this first-order structure. Specifically, the SRN model shows how broadly hierarchical patterns of behavior can emerge from a processing system that—because it is not constrained to represent only strictly hierarchical relationships—retains the flexibility to encode aspects of sequential action that violate strict hierarchy. Thus, no special mechanisms are needed to allow for interactions among levels of task structure.

In addition to capturing hierarchical relationships where they are critical to the guidance of behavior, the system's representational space retains sufficient flexibility that it can accommodate potentially complex interrelations among different operations. The representations it develops can respond simultaneously to the simi-

[4] This concern does not appear to run very deep because Cooper and Shallice (2006) themselves offered the appropriate response. They write that "one might envisage a system that maintains associations between higher-level representations of schemas and the hidden unit patterns that result in those schemas being performed. A supervisory system could then interface with the SRN model to yield controlled behavior (when required) by deliberately instantiating the hidden units with the corresponding activations . . . There is also a sense in which the instruction units already present in the SRN model do this for the two basic tasks of preparing tea and preparing coffee" (p. 905). This is, indeed, precisely the account we would offer, as indicated in Botvinick and Plaut (2004, p. 424). We assume that this form of interface would support recovery from errors, guided by the supervisory system.

larities and differences among sets of overlapping activities (Botvinick & Plaut, 2002). This property is important, given that complex patterns of overlap are characteristic of human behavioral repertoires (consider the relationships among the routines for spreading jam, spreading peanut butter, spreading sauerkraut on a hot dog, spreading icing on a cake, spreading wax on a floor, using a squeegee on a window, and raking the lawn).

It is worth remarking that it was considerations of this kind, and specifically the role of complex shared structure, that in the 1980s inspired David Rumelhart to abandon the symbolic approach to schema representation (Rumelhart & Ortony, 1977) in favor of a recurrent neural network account, laying some of the initial foundations for connectionist psychology (Rumelhart, Smolensky, McClelland, & Hinton, 1986). As noted in Botvinick and Plaut (2004), the SRN account we have proposed builds rather directly on this pioneering work, as does recent work in concept representation, another area in which the schema construct has a long history (Rogers & McClelland, 2004).

Conclusions and Directions for Future Work

In the present reply to Cooper and Shallice (2006), we answer a set of specific points from that critique and also examine some fundamental theoretical issues brought to the surface by the contrast between our two models. In revisiting the observations reported by Cooper and Shallice (2006) concerning the behavior of the SRN model, we argue that in almost all cases, they stem from incidental implementational choices rather than from core theoretical commitments. Making this point allows us to spell out our own assumptions concerning the patterns of experience on which learning of sequential action is based, which center on the concept of a representative sample. Next, we consider the claim from Cooper and Shallice (2006) that routine sequential action must rest on strictly hierarchical trees of localist task and subtask representations. This provides the opportunity to make explicit our alternative assumption of a less constrained representational space, which can accommodate quasi-hierarchical structure and simultaneously support information sharing and context sensitivity. Finally, discussing the apparent claim from Cooper and Shallice (2006) that habits and goal-directed actions must rely on isomorphic representations, we point to theoretical motivations for an alternative view according to which the two modes of action exploit different kinds of representation, much as different modes of spatial navigation are thought to depend on different kinds of spatial representation.

In offering the present reply to Cooper and Shallice (2006), our goal is to lay out some important theoretical alternatives rather than to present a conclusive argument. Although we bring empirical data to bear wherever possible, many of our assertions, like many of those made by Cooper and Shallice (2006), rest on logical considerations or appeals to common sense. There is a great need for additional empirical data to guide decisions among (or displace) the theoretical alternatives we lay out. Such data is challenging to generate, given the intrinsic complexity of the behavioral domain and the difficulty of distinguishing routine habit-driven action from more deliberate goal-driven behavior. Botvinick and Bylsma (2005) reported one empirical finding, concerning the impact of interruptions on sequence errors, that is directly relevant to the contrast between the SRN and IAN models, and it is disappointing that Cooper and Shallice (2006) did not address this in the terms directly linked to their computational model. On a broader level, we agree with Cooper and Shallice (2006) that the behavioral and neuroscientific work reported by Balleine and colleagues (e.g., Balleine & Dickinson, 1998) has provided some important empirical points of reference. The computational investigations reported by Daw et al. (2005), addressing that empirical research, also strike us as providing a useful new point of reference for the ongoing development of theories of routine sequential action.

References

Agre, P. E. (1988). *The dynamic structure of everyday life* (Tech. Rep. No. 1085). Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Aldridge, W. J., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: A "natural action" approach to movement sequence. *Journal of Neuroscience, 18,* 2777–2787.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science, 26,* 39–83.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94,* 192–210.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

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.

Avrahami, J., & Kareev, Y. (1994). The emergence of events. *Cognition, 53,* 239–261.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37,* 407–419.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications, 13,* 343–379.

Barto, A. G., Singh, S., & Chentanez, N. (2005). *Intrinsically motivated learning of hierarchical collections of skills.* Paper presented at the Proceedings of the Third International Conference on Developmental Learning (ICDL 2004), La Jolla, CA.

Botvinick, M. (in press). Hierarchical structure in behavior and in the brain: A model of Fuster's hierarchy. *Proceedings of the Royal Academy of London: Series B.*

Botvinick, M., & Bylsma, L. M. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin and Review, 12,* 1011–1017.

Botvinick, M., Bylsma, L. M., Buxbaum, L. J., & Jax, S. A. (in press). Plan-dependent distractor effects in object-directed action: Empirical observations and a computational account. *Journal of Experimental Psychology: Human Perception and Performance.*

Botvinick, M., & Plaut, D. C. (2002). Representing task context: Proposals based on a connectionist model of action. *Psychological Research, 66,* 298–311.

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.

Buxbaum, L. J., Schwartz, M. F., & Montgomery, M. W. (1998). Ideational apraxia and naturalistic action. *Cognitive Neuropsychology, 15,* 617–643.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing.* Cambridge, MA: MIT Press.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute

to planning and prospective memory. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications.* Hillsdale, NJ: Erlbaum.

Cooper, R. P. (2003). Mechanisms for the generation and regulation of sequential behavior. *Philosophical Psychology, 16,* 389–416.

Cooper, R. P. (in press). Tool use and related errors in ideational apraxia: The quantitative simulation of patient error profiles. *Cortex.*

Cooper, R. P., Schwartz, M. F., Yule, P. G., & Shallice, T. (2005). The simulation of action disorganization in complex activities of daily living. *Cognitive Neuropsychology, 22,* 959–1004.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology, 17,* 297–338.

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

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 1355–1367.

Dickinson, A. (1985). Actions and habits: The development of behavioral autonomy. *Philosophical Transactions of the Royal Society (London), Series B, 308,* 67–78.

Eck, D., & Schmidhuber, J. (2002). Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In H. Bourlard (Ed.), *Neural networks for signal processing XII: Proceedings of the 2002 Institute of Electrical and Electronics Engineers (IEEE) workshop.* New York: IEEE.

Forde, E. M. E., & Humphreys, G. W. (2002). The cognitive processes underpinning everyday actions. *Neurocase, 8,* 59–60.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences, 3,* 128–135.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe.* Philadelphia: Lippincott-Raven.

Grafman, J. (1995). Similarities and distinctions among current models of prefrontal cortical functions. In J. Grafman, K. J. Holyoak, & F. Boller (Eds.), *Structure and functions of the human prefrontal cortex* (pp. 337–368). New York: New York Academy of Sciences.

Hartley, T., Maguire, E. A., Spiers, H. J., & Burgess, N. (2003). The well-worn route and the path less traveled: Distinct neural bases of route following and wayfinding in humans. *Neuron, 37,* 877–888.

James, W. (1890). *The principles of psychology.* New York: Holt.

John, B. E. (2003). Information processing and skilled behavior. In J. M. Carroll (Ed.), *HCI models, theories, and frameworks: Toward a multidisciplinary science.* Boston: Morgan Kaufman.

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 are there complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6,* 465–472.

McClelland, J. L., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior.* New York: Holt, Rinehart & Winston.

Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing. *Connection Science, 6,* 247–280.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88,* 1–14.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davison, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4, pp. 1–18). New York: Plenum.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Poldrack, R. A., Prabakharan, V., Seger, C., & Gabrieli, J. D. E. (1999). Striatal activation during cognitive skill learning. *Neuropsychology, 13,* 564–574.

Reason, J. T. (1990). *Human error.* Cambridge, England: Cambridge University Press.

Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science, 11,* 5–40.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition.* Cambridge, MA: MIT Press.

Rougier, N. P., Noell, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences, 102,* 7338–7343.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the Acquisition of Knowledge* (pp. 99–136). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the microstructure of cognition* (pp. 7–57). Cambridge, MA: MIT Press.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44,* 493–515.

Salinas, E. (2004). Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *Journal of Neuroscience, 24,* 1113–1118.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding.* Hillsdale, NJ: Erlbaum.

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

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence, 112,* 181–211.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19,* 181–189.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127,* 3–21.

Zhang, W., Samaras, D., Yang, H., & Zelinsky, G. (2005). A computational model of eye movements during object class detection. *Neural Information Processing Systems (NIPS*). For more information see http://www.nips.cc

Appendix

Details of the Simulation Briefly Described in *Interchangeable Subsequences*

Task and Representations

As explained in the main text, the task domain included 10 sequences, each composed of four steps. The model's universe was assumed to include seven objects (denoted a_1, a_2, a_3, a_4, a_5, b_1, and b_2) only one of which could be viewed at a time. The universe was also assumed to contain 14 actions: 1 search action for each object (e.g., *fixate*-a_1) and 1 manipulative action appropriate to each object (e.g., *use*-a_1). Each target sequence began on Step 1 with an object $a_i \in \{a_1, a_2, a_3, a_4, a_5\}$ as the viewed object and the target response being *use*-a_i. On Step 2, the input remained unchanged, and the target action became *fixate*-$b_j \in \{b_1, b_2\}$. On Step 3, b_j became the viewed input, and the target action changed to *use*-b_j. On Step 4, with the viewed input remaining b_j the target output became a_i, redirecting the system back toward the object with which the sequence began. Crossing all choices of *a* with all choices of *b* yielded the 10 target sequences. As explained under *Results and Discussion*, a second version of the task assumed that on the first step an additional input was provided indicating, as if via peripheral vision, whether object b_1 or b_2 was present in the environment.

Model Architecture

The model was identical to the one used in Simulation 1 of Botvinick and Plaut (2004), except with regard to the following details. The model included 7 input units, each representing a single visual object. To represent the currently viewed object, on each step of processing 1 of these units was set to an activation level of 1 whereas the others were set to 0. Unlike the Botvinick and Plaut (2004) model, the present network did not include input units representing held objects. The output layer contained 14 units, each representing 1 of the actions identified above. The hidden layer contained 100 units.

To simulate the alternative version of the task described above and under *Results and Discussion*, an extended version of the model included an additional 2 "peripheral vision" units, allowing it to receive information, on the first step of the task, as to whether b_1 or b_2 was present in the environment. In sequences involving b_1, one of these input units was activated on the first step of processing, and on b_2 trials the other unit was activated. On subsequent trials, both units were inactive.

Training and Testing

The model was trained and tested following the procedure used in Botvinick and Plaut (2004). In a first simulation, the model was trained on all 10 basic target sequences, and the model was then tested on all 10 sequences. In a second simulation, all but 1 sequence ($a_1 \rightarrow b_2 \rightarrow a_1$) were presented during training, and the model was tested on the omitted sequence. In a final simulation, the extended task was used, again omitting and testing ($a_1 \rightarrow b_2 \rightarrow a_1$). In the latter two simulations, the sequence $a_1 \rightarrow b_1 \rightarrow a_1$ was presented twice as often as other sequences during training, to assure that the frequency of all *a* fillers was matched. In all simulations, the duration of training was 2,500 trials. To establish replicability, each simulation was repeated with each of 10 sets of random initial weights.

Results

When trained on all 10 sequences, the model successfully reproduced each sequence at test. On Step 2, *fixate*-b_1 and *fixate*-b_2 output units both assumed activation values of approximately 0.5, reflecting the uncertainty associated with this point in the sequence. More important was the question of whether the model would produce $a_1 \rightarrow b_2 \rightarrow a_1$ when this was omitted during training. Of particular interest was the model's behavior at two specific steps following initial presentation of a_1. First, we considered whether, on Step 2, the model would activate *fixate*-b_2 as well as *fixate*-b_1. Second, we considered whether, on Step 4, the model would most strongly activate *fixate*-a_1. Together, these behaviors would indicate that the model had inferred that the subsequences *fixate*-$b_1 \rightarrow$ *use*-b_1 and *fixate*-$b_2 \rightarrow$ *use*-b_2 are interchangeable.

In the first simulation in which generalization was tested, to our initial surprise the model did not show the first form of generalization. On Step 2, only *fixate*-b_1 was significantly activated. However, there turned out to be a sensible reason for this. Note that during training the model observed only b_1—never b_2—employed in the a_1 context without any explanation for this bias. This would be analogous to observing sugar added from a packet—and not from a sugar bowl—over thousands of witnessed executions of coffee making. Clearly, on the basis of this experience, it would be quite reasonable to infer that coffee making forbids the use of a sugar bowl. However, consider the same scenario if it were also known that, in each witnessed instance of coffee making, only sugar packets were available. This would provide an explanation for the failure to observe use of a sugar bowl, making it more reasonable to infer that sugar bowl and sugar packet use are interchangeable in the coffee making context, as elsewhere. To impose an analogous situation in our model, we included the peripheral vision units described above. These provided information at the outset of each trial as to whether b_1 or b_2 was present in the environment. When trained on this modified version of the task, the model showed both forms of generalization predicted, including correct responses on Step 4, on all 10 simulation runs. Presented initially with a_1 as the viewed object and peripheral vision input indicating the presence of b_2 in the environment, the model consistently generated the doubly novel sequence *use*-$a_1 \rightarrow$ *fixate*-$b_2 \rightarrow$ *use*-$b_2 \rightarrow$ *fixate*-a_1. This result contradicts the assertion in Cooper and Shallice (2006) that the SRN model has "no way of knowing how to preserve context information (e.g., whether it is making tea or coffee) across a subtask (e.g., adding sugar) unless it has received explicit training on that variant of the task" (p. 899).

Critically, as noted in the main text, the results differed when a smaller range of contexts was included in training. Specifically, when the training set included only the sequences $a_1 \rightarrow b_1 \rightarrow a_1$, $a_2 \rightarrow b_1 \rightarrow a_2$, and $a_2 \rightarrow b_2 \rightarrow a_2$, and the model was again tested for production of $a_1 \rightarrow b_2 \rightarrow a_1$, the model never correctly generalized on Step 2, even when the peripheral vision units were included.

Received February 3, 2006
Revision received April 20, 2006
Accepted April 24, 2006 ■

Postscript: The Way Forward

Matthew M. Botvinick
University of Pennsylvania

David C. Plaut
Carnegie Mellon University

In our view, the reply by Cooper and Shallice (2006) left our original responses more or less intact. Cooper and Shallice disparaged our account of anticipation errors by arguing that it serves only to "highlight the importance of the training set in shaping the [SRN] model's errors" (p. 892). However, as we argued earlier, this link between errors and prior experience may be viewed as a strength of the SRN model rather than a weakness. They dismissed the idea of selecting among actions probabilistically in the SRN model (an idea they earlier seemed to promote) on the basis of the claim that such a selection procedure cannot be biased by previous intentions. However, it is easy to see how the setting of intentions could be encoded in the context layer, leading to amplification of intended action outputs and the suppression of competitors. They rejected our demonstration that the SRN model can, under appropriate circumstances, deal appropriately with novel initial conditions, but they did this by resorting to the confusing argument that the model's correct performance must be understood as an error. They rejected our demonstration of object substitution errors in the SRN with the assertion that this class of error is defined in earlier work in terms of the actions following the actual incorrect use of an object, but we fail to find this kind of definition in the sources they cite (Reason, 1984; Schwartz et al., 1998). They responded to our analysis of the role of *goal nodes* in the IAN model with the assertion that this mechanism "captures the fact that some schemas have a common purpose" (p. 894). However, at the level of function (rather than description), goal nodes serve only to enforce the selection of one among a set of competing schemas. This functionality involves no direct reference to purpose or goal. Certainly, in human behavior, interchangeable action sequences often share a common goal. However, in the IAN model, this fact only informs the way that the network is wired up by the modeler. It is in no way intrinsically "captured" by the goal node mechanism.

Clearly, the present exchange will leave a range of questions unresolved. However, what we are left with is far from a hopeless impasse. Instead, the debate has made clear what measures might be taken to shed further light on the mechanisms underlying routine sequential behavior. Clearly, one road forward involves implementing and evaluating the various refinements and extensions that have been proposed for both the SRN and IAN models. With regard to the SRN model, it may be particularly informative to develop further the simulation we offered to address the topic of interchangeable subsequences, considering more fully the problem of implicit negative evidence (the problem that led us to include "peripheral vision" inputs to the framework). With regard to the IAN model, what seems most urgently needed is an explicit, implemented account of how the goal-directed action system (or supervisory system) trains up the habit system (contention scheduling system). Here and in general, the present exchange has strongly highlighted the degree to which any account of the habit system will interact with how the goal-directed system is understood. Many of the points we have debated have turned out to depend on what assumptions are made about the goal-directed system and about how it interfaces with the habit system. In view of this, what seems necessary is to place our present accounts of routine sequential action within a larger computational framework that explicitly addresses both systems and their interrelations. We agree with Cooper and Shallice that such an enterprise might best concentrate on instances of behavior that depend critically on a collaboration between the two systems, such as error correction or coping with novel and unexpected environmental contingencies encountered during routine behavior. Building a larger theoretical framework, capable of dealing with such areas of behavior, will require innovative experimental research, because there are at present frustratingly few meaningful empirical benchmarks to constrain new modeling.

References

Cooper, R. P., & Shallice, T. (2000). Structured representations in the control of behavior cannot be so easily dismissed: A reply to Botvinick and Plaut (2006). *Psychological Review, 113,* 929–931.

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.

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.