Still at the Choice-Point

Action Selection and Initiation in Instrumental Conditioning

BERNARD W. BALLEINE AND SEAN B. OSTLUND

Department of Psychology and the Brain Research Institute, University of California, Los Angeles, California, USA

ABSTRACT: Contrary to classic stimulus–response (S-R) theory, recent evidence suggests that, in instrumental conditioning, rats encode the relationship between their actions and the specific consequences that these actions produce. It has remained unclear, however, how encoding this relationship acts to control instrumental performance. Although S-R theories were able to give a clear account of how learning translates into performance, the argument that instrumental learning constitutes the acquisition of information of the form “response R leads to outcome O” does not directly imply a particular performance rule or policy; this information can be used both to perform R and to avoid performing R. Recognition of this problem has forced the development of accounts that allow the O and stimuli that predict the O (i.e., S-O) to play a role in the initiation of specific Rs. In recent experiments, we have used a variety of behavioral procedures in an attempt to isolate the processes that contribute to instrumental performance, including outcome devaluation, reinstatement, and Pavlovian–instrumental transfer. Our results, particularly from experiments assessing outcome–selective reinstatement, suggest that both “feed-forward” (O-R) and “feed-back” (R-O) associations are critical and that although the former appear to be important to response selection, the latter—together with processes that determine outcome value—mediate response initiation. We discuss a conceptual model that integrates these processes and its neural implementation.

KEYWORDS: goal-directed action; reward; habit; reinforcement; associative learning

INTRODUCTION

Whether framed in computational or, more explicitly, in psychological terms, contemporary theories of decision making commonly rely on the cognitive
or executive capacities of the agent to explain choice.\textsuperscript{1–6} As a consequence, these positions can slip into indeterminacy and it is easy to see why. Any information that takes the form “action A leads to outcome O” can be used both to perform A and to avoid performing A. It is simply not possible purely on the basis of the information presented to an agent, such as “A leads to O” and “B leads to P,” to predict whether they will choose A or B because, whether derived from perceptual, cognitive, social, or fictive sources, information alone is not sufficient to determine a course of action. This criticism of cognitive, information-based theories of action was recognized early in analyses of animal behavior. The classic critique in that context was Guthrie’s\textsuperscript{7} jibe at the cognitive behaviorism of Edwin Tolman,\textsuperscript{8} particularly the latter’s contention that the performance of a rat learning to traverse a maze to find food was a matter of acquiring a belief about “what (action) leads to what (outcome).” As Guthrie\textsuperscript{7} put it, on this view “the rat is left buried in thought” (p. 172). Merely believing that, say, “turning left at the choice point is necessary to get food” does not entail turning left.

Guthrie himself favored an account of animal action based on the formation of sensorimotor, so-called stimulus–response (S-R), associations—a theory that rendered actions homologous to reflexes and that explained performance by confounding action selection and initiation within the function of the stimulus. The most influential version of this account was later developed by Hull,\textsuperscript{9} who proposed that S-R associations are strengthened by reinforcement; that is, an association between the situational stimuli (S) and a response (R) is strengthened when R is followed by a reinforcing event (such as food) thereby accounting for the observation that R becomes more probable in S. However, as has been well documented in the past,\textsuperscript{10,11} this account means to claim that animals do not encode the consequences associated with their actions; that is, the reinforcer or outcome contingent on the performance of R does not itself form a part of the associative structure controlling the performance of that R. Indeed, it was recognition of this fact that produced some of the critical experimental tests of S-R theory, notably the \textit{outcome devaluation test.}\textsuperscript{11}

The outcome devaluation test, conducted after training and so after the formation of any S-R connection has been made, involves changing the value of the instrumental outcome using any of a variety of motivational manipulations, such as taste aversion learning\textsuperscript{10,12} or specific satiety,\textsuperscript{13,14} after which the tendency of the rat to press the lever is assessed in extinction, that is, in a test in which no outcomes are delivered. Performance of the devalued action is compared either against that of a nondevalued control or, in a choice situation, against the performance of another nondevalued action. S-R theories of instrumental performance predict that, because of the S-R association established during training, the presence of the training S guarantees that R will be performed on test irrespective of the change in value of the training outcome. In direct conflict with this prediction, however, numerous experiments have found that choice between actions respects the current value of the specific
outcome associated with an action and not the presence or absence of some eliciting stimulus or other.12,15,16

These findings suggest that choice is at least partly determined by integrating action–outcome associations with outcome values. Nevertheless, although this kind of account can provide a good basis for action selection, it is not immediately clear how it determines performance; that is, how does believing that “A leads to O” and “O is valuable” induce the agent to perform A? One can imagine that knowing A→O and B→P and that O has greater value than P could lead one generally to prefer A over B. But does that necessarily result in greater performance of A? Is it sufficient that Jack prefers coffee to tea and knows their different means of production for Jack to make coffee? Although, on this account, something further than selection is required for action initiation, the relationship between selection and initiation is unexplored and remains implicit in the description of performance. Is there any alternative?

TWO-PROCESS THEORY

Various alternatives to the strict S-R account described earlier have been developed that, given certain assumptions, can predict the effects of outcome devaluation on performance without resorting to action–outcome learning. The most influential of these has been two-process theory. There have been two variants of this theory, a motivational version17 and an expectancy version.18 Both versions propose that instrumental learning is fundamentally S-R and that performance depends on the influence of predictions of the outcome (O) based on certain state or situational cues (i.e., S-O), but differ in the details of the influence of these predictions: on the motivational account, S-O associations modulate performance elicited by the S-R association, whereas on the expectancy account, the expectancy of O forms a component of the S with which the R becomes associated, that is, S-(O)-R.

The standard actor–critic formulation of reinforcement learning models has most in common with the first of these alternatives.19,20 In that model, learning, that is, the acquisition of policies, is essentially a matter of forming (internal and/or external) state–response associations strengthened by the (reinforcing) feedback derived from the value of the state to which the response allows a transition. States have values based on the outcomes that they predict. Hence, together, states both motivate actions and reinforce them. Unfortunately, both the motivational version of two-process theory and its actor–critic implementation have difficulties explaining choice performance in an outcome devaluation test. Simply put, unless differential predictors can be established, a state, Sₙ, that is trained with two actions, Sₙ→R₁; Sₙ→R₂, for distinct but equally valued outcomes cannot explain the well-documented changes in choice performance that ensue when one or other outcome is subsequently devalued; the state predicts both a valued and a devalued outcome and cannot proscribe
which of the two actions associated with the state should be performed. In
one example, 21 hungry rats were trained to push a pole in one direction to
get grain pellets and in the opposite direction to get sucrose. In this situation,
both policies should be equally preferred; both actions are equally associated
with all potential state cues, and the value of the state transitions, or outcomes
of each action, are equal. Nevertheless, after outcome devaluation, animals
immediately, that is, without further training, reduce their tendency to push in
the direction that previously resulted in the now devalued outcome.

The alternative, expectancy formulation of two-process theory has had some-
what greater success because it allows the assumption that policies are deter-
dined by state–state transitions. For example, this account would propose that
the policy R1 under S1 that results in a transition to S1 is controlled by S1S1
(i.e., S1S1–R1), and of R2 resulting in a transition to S1 is controlled by S1S2
(i.e., S1S2–R2). If changing the value of S1 relative to S2 is assumed to imme-
diately alter the ability of S1 to retrieve S2 relative to S1, then a basis might
be established for predicting the relative change in performance induced by
deviation in the choice test in the study of Dickinson et al. among others. 12, 21
In associative terms, this is the equivalent of the expectancy of an outcome
O that is associated with a stimulus, that is, S-(O), controlling the production
of the response, that is, S-(O)-R, an association that is strengthened by the
reinforcement of R in the presence of S-(O) in a manner that accords with S-R
theory. Psychologically, this account can explain the relative reduction in per-
performance of an action associated with a devalued outcome either by supposing
that the expectancy of an outcome is suppressed (or relatively so) when that
outcome is (relatively) devalued, or that by retrieving O, S also retrieves an
aversive emotional state reducing the performance of the response accordingly.

There is evidence that accords with this version of two-process theory; when
the expected outcome and the delivered outcome are the same, animals learn
more rapidly than when they are different. For example, consider the experi-
mental design described in Design 1 taken from an experiment conducted by
Trapold and Overmier. 18 In this study, rats were trained first to predict two
distinct outcomes (O1 and O2) on the basis of two stimuli (S1 and S2) after
which the stimuli were used to discriminate which of the two responses (R1 and
R2) would be reinforced. Two groups were used in this discrimination phase: one
in which the predicted and earned outcomes were congruent and one in which
they were incongruent (see Design 1).

Congruent: [S1-O1; S2-O2] then [(S1: R1-O1; R2-); (S2: R1-; R2-O2)]
Incongruent: [S1-O1; S2-O2] then [(S1: R1-O2; R2-); (S2: R1-; R2-O1]

(1)

Trapold and Overmier 18 found that, when the outcome expected on the basis
of S1 was also earned by R1, rats learned the discrimination more rapidly
than when it differed from the outcome earned by R1; that is, in accord with
predictions from the S-(O)-R account, Group Congruent learned faster than Group Incongruent. Similarly, it was subsequently found, as described in Design 2, that when rats were trained to perform two actions for distinct outcomes, presentation of a stimulus associated with one or the other outcome elevated the performance of the action trained with that outcome more than performance of the other action, an instance of an outcome selective form of a phenomenon referred to as Pavlovian–instrumental transfer.\(^{22}\)

\[
\text{If } [S1-O1, S2-O2] \text{ and } [R1-O1, R2-O2] \\
\text{then } [(S1: R1 > R2) \text{ and } (S2: R1 < R2)]
\]

Thus, again in line with the S-(O)-R account, increasing the expectancy of O1 by presenting S1 was found to increase the performance of the R trained with O1, in this case R1 relative to R2; indeed performance of R2 did not differ from a baseline (no stimulus) period.

**HIERARCHICAL THEORY**

Despite this clear evidence in favor of the S-(O)-R account, there is even more compelling evidence against it as a general explanation of instrumental performance. First and foremost, studies assessing the neural bases of instrumental conditioning have found evidence that distinct structures mediate outcome devaluation and the influence of Pavlovian S-O associations on instrumental performance. For example, Corbit _et al._\(^{23}\) compared the effects of lesions of distinct nuclei in the ventral striatum, the core and shell subregions of the nucleus accumbens, on outcome devaluation effects induced by specific satiation and on the Pavlovian–instrumental transfer effect reported by Colwill and Rescorla.\(^{22}\) Sham lesioned control subjects showed both devaluation and selective transfer effects. In contrast, and against predictions from the S-(O)-R account, lesions of the accumbens shell were found to abolish the excitatory effect of reward-related cues on instrumental performance without affecting outcome devaluation. Furthermore, although lesions of the accumbens core were found to abolish outcome devaluation, they had no effect on Pavlovian–instrumental transfer. This evidence that devaluation and transfer effects can be doubly dissociated at the level of the nucleus accumbens makes the argument that the explanation of one effect should be made in terms of the other particularly difficult to maintain (see also Ref. 24).

Furthermore, as described earlier, devaluation effects using a bidirectional manipulandum can only be explained by two-process theory if the expectancy based on state–state transitions argued to control performance is suppressed by outcome devaluation. Several studies have found that this is not so. For example, using a design similar to that described earlier by Trapold and Overmier, Rescorla\(^{21}\) first trained explicit S1-O1, S2-O2 expectancies and then trained the rats to perform two actions, R1 and R2, one in the presence of each S, such
that the reward delivered was incongruent with that predicted by S, that is, S1: R1-O1; S2: R2-O1. During this training, S1-O1 and S2-O2 continued to be presented intermixed with the instrumental contingencies. He then devalued O1, using a taste aversion procedure, and assessed the choice between R1 and R2 in the presence of both S1 and S2. On the S-(O)-R account, devaluation should have suppressed the control of S1 over R1 resulting in R1 < R2 on test. In direct contradiction of this prediction, Rescorla found that, in the presence of S1S2, R2 was performed less than R1; that is, the performance of the rats in this experiment respected the expectancy of the outcome based on their actions rather than that based on S1 and S2. Similarly, outcome devaluation after training on a biconditional discrimination, should not, on the two-process account, be predicted to generate differential performance. Rescorla26 reports the results of a study along these lines using a design similar to that described in Design 3, in which two responses were trained for different outcomes in each of two stimuli after which O1 was devalued (i.e., O1-):

\[
\begin{array}{c|c|c}
S1: & R1-O1, R2-O2 & O1- \quad | \quad S1: R1 < R2 \\
S2: & R1-O2, R2-O1 & S2: R1 > R2 \\
\end{array}
\]

Although, on the S(O)-R account both S1 and S2 should be equally impaired in their control of R1 and R2 after the devaluation of O1, a choice test between R1 and R2 found that, under S1, R1 was reduced relative to R2 whereas, under S2, R2 was reduced relative to R1.

Together these kinds of findings have encouraged the development of a hierarchical theory of instrumental conditioning in which discriminative stimuli, such as S1 and S2, are proposed to control the association between response and outcome by reducing the threshold required for activation of the outcome representation by the action12,26 (see Fig. 1A). On this view, instrumental performance is not a simple matter of binary S-O, S-R, or R-O associations but reflects the hierarchical control of S on the R-O association; whenever the controlling S is present and the response is available, the S-(R-O) structure ensures that the instrumental response will be performed.

Although the hierarchical account is consistent with much of the data, recent tests of predictions derived from this position have raised doubts about its veracity. One particularly difficult set of observations for this view has come from what may be called “component discriminations,” which can be used to model the discriminative properties of free operant schedules (see Design 4). In free operant conditioning, training is typically conducted in sessions in which only one response manipulandum is available at any one time. This arrangement ensures that the outcome earned by performing that response maintains a consistent relationship both as a consequence of the response but also as an antecedent or discriminative stimulus for the next response—refer to Design 4(a). It is, however, possible to arrange a situation in which the outcome that serves as the antecedent differs from that earned as a consequence
FIGURE 1. (A) The hierarchical model of instrumental conditioning according to which discriminative stimuli lower the threshold on the outcome representation increasing its activation by the associated response to generate performance. (B) Predictions from the hierarchical theory for the congruent and incongruent component discriminations described in the text. When the outcome used as the discriminative stimulus is congruent with that earned by the response, the discrimination is readily solved because O1 lowers the threshold of O1 resulting in R1 > R2, and O2 lowers the threshold of O2 resulting in R2 > R1. When the outcome used as the discriminative stimulus differs from that earned by the response, however, O1 will lower the threshold on O2, activation of which will lower the threshold on O1 making both R1 and R2 equally likely and the discrimination relatively more difficult. (C) Data collected at UCLA by Sanne de Wit and Sean Ostlund comparing the congruent and incongruent discriminations that replicate the finding that, in contrast to the predictions from hierarchical theory, the incongruent discrimination is acquired more rapidly than the congruent (the asterisk signals that $P < 0.05$ for the groups’ comparison).

of responding, for example, in a situation in which two responses (R1 and R2) that earn different outcomes (O1 and O2) are performed in strict alternation—refer to Design 4(b).

(a) R1-O1-R1-O1-R1 . . . etc.
R2-O2-R2-O2-R2 . . . etc. (4)

(b) R1-O1-R2-O2-R1 . . . etc.

As illustrated in Design 5, component discriminations establish the components of the two kinds of free-operant schedule described in Design 4 as
discrete problems; that is, in the congruent case, the delivery of O1 signals that R1, not R2, will deliver O1 and O2 signals that R2, not R1, will deliver O2 whereas, in the incongruent case, O1 signals that R1, not R2, will deliver O2 and O2 signals that R2, not R1, will deliver O1.

\[
\text{Congruent: [O1: R1-O1, R2-] and [O2: R2-O2, R1-]}
\]
\[
\text{Incongruent: [O1: R1-O2, R2-] and [O2: R2-O1, R1-]}
\]

Although both of these discriminations model components of free-operant schedules, hierarchical theory predicts that the rate of acquisition of the congruent and incongruent discriminations will differ considerably (see Fig. 1B). Hierarchical theory has no difficulty predicting successful discrimination performance in the case of the congruent discrimination; delivery of O1 reduces the threshold on the R1-O1 association resulting in the performance of R1 whereas delivery of O2 reduces the threshold of the R2-O2 association resulting in the performance of R2. Hence, delivery of O1 will result in R1 > R2, whereas O2 will result in R2 > R1. In contrast, hierarchical theory predicts that the incongruent discrimination should be much harder to solve. Consider what should happen following the delivery of O1. Because O1 signals that R1 will be reinforced by O2, the delivery of O1 should allow R1 to more readily activate O2. On this account, however, activation of O2 should in turn lower the threshold for R2 to activate O1. As a consequence, hierarchical theory makes the prediction that the delivery of O1 (or O2) should result in an increased tendency to perform both R1 and R2, thereby making this discrimination much more difficult and perhaps unsolvable.

Experiments assessing the rate of discrimination learning in this situation have not supported hierarchical theory. For example, Dickinson and de Wit\(^\text{27}\) trained rats on two different component-discrimination problems, one in which the delivery of a drop of sucrose signaled that a left lever press response would deliver a drop of sucrose and a right lever press nothing and another on which a food pellet signaled that a right lever press response would deliver a pellet and a left lever press nothing (the responses were counterbalanced across subjects). In this case, the discriminative (antecedent) cue was congruent with the outcome (or consequence) of responding. A second group of rats was also trained on two component discriminations, but in this case the antecedents and consequences were incongruent; that is, the sucrose signaled that a left lever press would deliver a pellet and a right lever press nothing, and a pellet signaled that a right lever press would deliver the sucrose and a left lever press nothing (again, counterbalanced). It is this latter pair of discriminations that hierarchical theory predicts should be relatively difficult to solve because the delivery of either outcome should provoke the tendency to press both the left and right levers, thus leading to response conflict.

In fact, despite this clear prediction of hierarchical theory, it was not upheld in the results of this experiment. Indeed, Dickinson and de Wit\(^\text{27}\) reported
precisely the opposite result: they found that the incongruent discrimination was acquired faster than the congruent one. This surprising finding has been replicated in our lab. During a visit to UCLA by Sanne de Wit, we collaborated on an experiment in the course of which we replicated the procedure described by Dickinson and de Wit and found, again, that the incongruent discrimination was acquired faster than the congruent one (see Fig. 1C). There are at least two reasons why this might be so in this case. First, it is possible that the sucrose (or pellet) delivery produced a very short-term satiety or negative priming effect reducing the processing of the subsequent sucrose (or pellet) delivery, thereby reducing the rate of acquisition of the lever press sucrose (or pellet) association. Alternatively, although the presentation of the outcomes that served as the discriminative stimuli were not explicitly paired with any stimulus, they were, of course, presented in a specific set of contextual cues, the same context in which the lever press response was performed and paired with the outcome. This sets up a kind of relative-validity problem, wherein the context–outcome (C-O) association competes with the response–outcome (R-O) association by virtue of being present when the response is performed (C-O, C+R-O). However, this competition between context and response must be at last partially localized to individual trials; the context is an equally good predictor of the outcomes in the incongruent discrimination. This second account reduces, therefore, to a variant of the negative priming account in which the C-D association interferes with processing the R-O association.

Finally, it is worth pointing out that the ability of the rats to solve the incongruent discrimination suggests that they have mechanisms able to resolve the response conflict inherent in this discrimination problem. Studies assessing the neural basis of response conflict in humans using tasks such as the Stroop, Simon, or flanker tasks suggest that a region of the dorsal medial prefrontal cortex plays a critical role in this capacity.28 In a recent study, we have found evidence that the same is true of rats when solving the incongruent discrimination.29 Indeed, temporary inactivation of the dorsomedial prefrontal cortex (using the GABA-A receptor agonist muscimol) not only made this discrimination much more difficult, it rendered performance inferior to the control discrimination, not by reducing performance of the correct response in the discrimination but by increasing performance of the incorrect response. It is interesting to note, therefore, that the result predicted by hierarchical theory for normal rats does in fact emerge but only in rats with a deficit in medial prefrontal cortical function.

Generally, however, it should be clear that hierarchical theory, as currently formulated, cannot account for these data. Unless a principled argument can be developed that allows the sensory and expectancy properties of the outcome to be represented differently, this theory is forced to predict that the ability of rats to solve the incongruent discrimination should be reduced relatively to the incongruent discrimination.
AN ALTERNATIVE TWO-PROCESS ACCOUNT

Although two-process theory provides a relatively simple explanation for certain features of instrumental performance, the prior analysis suggests that variants of this theory, whether framed in S-(O)-R or S-(R-O) terms, fail to describe accurately the nature of the discriminative properties of the outcome in response selection and, as a consequence, do not generate an adequate account of instrumental performance. This should not be taken to imply that all forms of two-process theory are untenable. Indeed, recent evidence from our lab investigating selective reinstatement effects has provided evidence for just such an alternative two-process account.32

It has been well documented, using both Pavlovian30 and instrumental conditioning procedures,31 that, after a period of extinction, brief reexposure to the outcome will reinstate performance of the conditioned response or instrumental action, respectively. We have recently been investigating the outcome specificity of instrumental reinstatement.32 In these experiments, rats were given free-operant training on two levers for distinct outcomes (R1- O1, R2-O2) after which both actions were subjected to a period of extinction (R1-, R2-). During a final test phase, the rats were given a choice test on the two levers, again in extinction, immediately after the delivery of one or other instrumental outcome, for example, O1: R1 versus R2. In several studies, we have found that outcome delivery almost completely reinstated responding on the action with which it was associated during training.

The similarity of this effect to the component discriminations above raises the possibility that the selective reinstatement effect is mediated by the animals’ learning, during free-operant training, that the outcome serves not only as a consequence but as an antecedent of the response with which it was paired, that is, O1:R1-O1. This is the same as the congruent discrimination arranged by Dickinson and de Wit.27 Of course, the finding that reinstatement is selective not only accords with this account based on the antecedent O-R association, it also accords with the account that the free outcome restores performance though retrieval of the R-O association, perhaps through a backward R-O association. To examine these distinct explanations for selective reinstatement, we assessed the response specificity of reinstatement in a group given congruent training, in which the outcome signaling a response was always the same as the one earned by that response (i.e., O1:R1-O1 & O2:R2-O2), and a group given incongruent training, in which the outcome signaling a response differed from the outcome earned by that response (i.e., O1:R1-O2 & R2:R2-O1; Ostlund and Balleine32). The O-R and R-O accounts of reinstatement make distinct predictions with regard to the responses that will be reinstated on these two schedules. If the R-O association mediates reinstatement, then the delivery of a specific outcome (e.g., O1) should
reinstate the same response (i.e., R1) in both groups. If, however, it is the antecedent O-R association that mediates the reinstatement effect then these two training procedures should generate opposing effects; the outcome should reinstate the response it signaled during training regardless of which response it followed. The results of this experiment, presented in Figure 2A, demonstrate that selective reinstatement effects are a product of the antecedent outcome–response association established during training; whereas O1 reinstated R1 in the congruent training group, it reinstated R2 after incongruent training. Of course, it is possible that animals in the incongruent condition failed to learn the specific R-O associations presented during training. In a subsequent outcome devaluation test, we found, against this claim, that both groups selectively suppressed performance of the action that had earned the now devalued outcome, indicating that both congruent and incongruent training had generated substantial R-O learning.

This experiment suggests that the antecedent O-R association is responsible for guiding the reinstatement of instrumental performance. Consistent with this claim, we have also found evidence that O1 will reinstate both actions equally when rats are trained with stable R-O, but unstable O-R, relationships, that is, one in which both R1 and R2 were preceded equally often by both O1 and O2 (see Fig. 2A). In this study, the stable R-O associations were maintained but were presented such that the opportunity to perform R1 to gain access to O1 and to perform R2 to gain O2 occurred equally often after both O1 and O2. Hence, although the consequent R-O relationships were consistent, the antecedent O-R associations were mixed such that O1 and O2 were equally often associated with both R1 and R2. If the O-R association mediates reinstatement, then the delivery of, say, O1 should have provoked reinstatement of both R1 and R2 and, indeed, this is what we found (see Fig. 2A).

These data make clear the potential for an alternative two-process account, not based on S-O and R-O processes, but based on the formation of O-R and R-O associations in the course of conditioning. It remains to be considered how these processes are formally related to each other, and, in what follows, we present both a model of this relationship and evidence supporting predictions derived from that model.

**S-R versus O-R LEARNING**

The suggestion that outcome–selective reinstatement is induced through the O-R association predicts, by analogy to other forms of S-R process, that reinstatement will not be affected if the reinstating outcome is devalued. We have assessed this prediction using a motivational shift from hunger to thirst and, more recently, using devaluation by specific satiety—see Design 6. In both cases, we found evidence suggesting that although the influence of the
FIGURE 2. (A) Choice performance during a reinstatement test conducted after various types of training and the subsequent extinction of R1 and R2. (B) Choice between two levers during a reinstatement test after training on R1-pellet and R2-sucrose. Prior to extinction and test, one of the two outcomes was devalued for one group of rats (Devalued Group) but not the other (Nondevalued Group). Note that although the overall rate of performance during the test was strongly affected by devaluation, the proportional increase in responding generated by the devalued outcome was very similar to that generated by the nondevalued outcome.
reinstating outcome on response selection does not depend on outcome value, the overall degree of reinstatement is diminished by these treatments.

\[
\begin{array}{c|c|c}
\text{R1-O1, R2-O2} & \text{O1-} & \text{R1, R2} \\
\hline
\text{O1-} & \text{R1, R2} & \text{O1: R1 vs. R2} \\
\text{(Non)} O2: R1 vs. R2
\end{array}
\] (6)

In the latter study, hungry rats were trained to press two levers, one for pellets and the other for a sucrose solution. Next, rats were allowed to consume freely one of the two outcomes for 1 h immediately before undergoing a reinstatement test. It is important to note that, after responding was extinguished on both levers, the Devalued Group was given noncontingent exposure to the outcome on which they were sated, whereas the Nonvalued Group received the other (valued) outcome. As predicted, whether devalued or nondevalued, outcome delivery resulted in similar shifts in subjects’ choice of the action that had earned the reinstating outcome relative to their baseline choice (see Fig. 2B).

Nevertheless, as is clear from this figure, and as we have found generally, the pattern of reinstatement that emerged also depended heavily on outcome value; whereas the Nonvalued Group performed substantially more responses for the reinstating outcome than at the extinction period, this elevation was strongly attenuated in the devalued group. Indeed, the reinstatement effect that emerges after devaluation is derived from the relative difference produced by a mild increase on the reinstated action and a strong decrease in responding on the other action (see Fig. 2B).

This pattern of results is in fact more generally consistent with the view that both O-R and R-O associations contribute to the overall pattern of responding during reinstatement. Although the O-R association is necessary to select the action associated with the outcome during training, the actual degree to which that association results in the initiation of the response depends on the R-O association and, hence, on the value of the associated outcome. This pattern of results, therefore, suggests two things. First, it suggests that the effect of the O-R association is, generally speaking, on response selection; it can induce a bias into choice performance that is multiplied by the R-O association to determine the overall rate of performance. Second, it suggests that if the O-R association is sufficiently strong, it can induce a degree of performance in and of itself and independently of the R-O association.

In fact, this latter suggestion is not new and has been advanced previously to explain the residual instrumental performance that is often observed in tests of outcome devaluation, although in that context it is usually thought to reflect the effect of more general S-R associations on performance rather than the specific O-R association proposed here. Nevertheless, there is no doubting that the O-R association, as proposed, has much in common with standard S-R theoretical accounts and, indeed, can be seen as identifying one among a large number of other potential stimuli that could exert control over instrumental performance,
FIGURE 3. An elemental view of the formation of selective O-R associations during the training of two instrumental actions for distinct outcomes allowing for the selection of one response over another. Elements specific to R1 and R2 and that allow the selection of one response over another are shown as “a” and “b,” whereas the other elements are common to both responses. Note that the reinforcement signal (+) functions to strengthen the connection between the specific stimulus element and the response.

particularly in the case of a single action. Of course, in the choice situation, the sources of differential response selection are more limited—the context or contextual elements being equally associated with both actions. Even so, in the choice situation, any stimulus that retrieves a specific outcome should, through the O-R association, be expected to select a specific action and, if owing to prior reinforcement (+) that association is sufficiently strong, to result in the initiation of that action irrespective of the value of the outcome associated with that action. This kind of structure is illustrated in Figure 3.

There is in fact evidence from a range of sources that supports this analysis. First, it is clear from a number of studies that the strength of S-R associations embedded in the instrumental training situation increases progressively with extended training. Early in training, outcome devaluation effects are much larger than after continued training and, indeed, as Adams has shown, with sufficient training the effects of devaluation can be abolished altogether. There is, therefore, good reason to believe that the amount of performance controlled by the O-R association increases with training as this association strengthens with each reinforced action and certainly sufficient evidence to propose that even when devaluation effects are quite strong, the O-R association may well be strong enough to induce the amount of reinstatement performance observed after devaluation, as illustrated in Figure 2B.
Furthermore, several studies have found that outcome devaluation does not necessarily affect the excitatory influence of stimuli associated with the instrumental outcome on instrumental performance in demonstrations of selective Pavlovian–instrumental transfer.\textsuperscript{35–37} Just as we have found in our assessment of selective reinstatement, devaluing the outcome predicted by a stimulus can leave the excitatory influence of that stimulus on the performance of actions associated with that outcome relatively unaffected. Interestingly, this effect appears to be influenced by the degree of training. Holland\textsuperscript{37} has reported that actions that have been given relatively little training are not much influenced by Pavlovian stimuli; much as one should propose for reinstatement, if the O-R association is relatively weak, both reinstatement and transfer should be correspondingly weak. As training progresses, however, the influence of Pavlovian cues and, potentially, the outcome itself on performance increases, presumably because the strength of the O-R association activated by the stimuli (or the outcome) increases with this training.

Generally, therefore, these results and, indeed, this overall analysis suggest that O-R associations have much in common with S-R learning and can be regarded as being acquired and strengthened in much the same way, that is, by contiguity.\textsuperscript{38,39}

**INTEGRATING O-R AND R-O ASSOCIATIONS: THE ASSOCIATIVE–CYBERNETIC MODEL**

Some time ago, Dickinson and Balleine\textsuperscript{40} advanced a general model of instrumental performance based on the interaction of S-R and R-O processes in what was referred to at the time as an associative–cybernetic model. Although this model was proposed in that paper as a vehicle for explaining differences between Pavlovian-conditioned responses and instrumental actions, particularly with respect to their sensitivity to degradation of the R-O contingency, it was also found that this model was able to offer a general account for the effects of overtraining on instrumental outcome devaluation and, more recently, has also been argued to have a degree of neural plausibility.\textsuperscript{41,42} We contend that this model also provides a good architecture for understanding the interaction of O-R and R-O processes in the selection and initiation of instrumental actions, both in single-action training and in a choice situation. A revised version of this model that captures the various aspects of O-R and R-O associative processes described earlier and their integration is presented in Figure 4A.

In this model, action selection is largely controlled by outcome–response learning and hence by outcome–retrieval. Action selection initiates a process of action evaluation through the R-O association; that is, the value of the action is estimated on the basis of the predicted reward value of the outcome that is contingent on that action. Finally, the action selection and
(A) S-R memory

\[ S^0 \rightarrow S^0 \rightarrow S^1 \rightarrow R^0 \rightarrow R^1 \rightarrow R^2 \rightarrow \ldots \]

(B) Associative memory

\[ \text{ACC} \quad \text{S}^0 \quad \text{Pl} \quad \text{O} \quad \text{DMS} \quad \text{R-O} \quad \text{VA} \quad \text{MD} \quad \text{MD} \quad \text{VA} \quad \text{SNe} \quad \text{VTASNe} \quad \text{VTASNe} \quad \text{S-R memory} \quad \text{Incentive memory} \]
evaluation processes combine to initiate an action. The key to performance of any action is that response production will occur when the response representation in the motor system is activated above a low fluctuating threshold. Hence, performance of an action could occur without training when the threshold of this unit is randomly low or when activation of the representation is driven supra-threshold through activation by the response selection and evaluation processes. These sources of performance will, individually, result in a relatively low probability of a response and so highly variable rates of performance, particularly early in training. The probability of responding and the rate of performance generally is increased when both the feed-forward response selection process and the feedback evaluative process activate motor output together, guaranteeing supra-threshold levels of activation. On this view, although noise can result in a low fluctuating probability that a response will occur, consistent execution requires the integration of O-R and R-O processes at motor output. The occurrence of the response can result in the occurrence of its contingent outcome (O), and, when the latter occurs, this activates the outcome representation in the S-R memory, the associative memory, and the reward memory. The latter two update the strength of the R-O association and allow estimation of outcome value. At the same time, outcome delivery also gives rise to a reinforcement signal that strengthens contiguously active stimulus and response representations in the S-R memory, including the specific outcome–response association, as long as the outcome is appetitive. An aversive outcome or an outcome that has been devalued will inhibit the production

FIGURE 4. (A) A version of the associative cybernetic model. This version includes three memory modules: an S-R memory, composed of various S-R associations including the O-R association; an associative memory, composed of the S-O and R-O associations underlying Pavlovian and instrumental conditioning; and a reward memory that encodes the affective (appetitive [AP] or aversive [AV]) and emotional value of the outcome and separating its rewarding (Rew) and reinforcing (+) functions. As with earlier versions, performance is a product of the dual influence of S-R and associative memory modules on the motor system. Here, the latter influence is gated by the incentive memory. (B) Neural implementation of the associative–cybernetic model. Associative memory for instrumental conditioning is shown linking associative regions of prefrontal cortex—prelimbic area (PL), anterior cingulated (ACC), medial agranular area (mA), and posterior dorsomedial striatum (DMS) with substantia nigra pars reticulata (SNr) output and feedback via mediodorsal thalamus (MD) modulated by dopaminergic input to the DMS from the ventral tegmental area (VTA). The S-R memory is composed of sensorimotor cortex (S-M), dorso-lateral striatum (DLS) pallidum (GPi), and thalamic feedback (via ventral anterior [VA] and the ventrolateral nuclei) again modulated by dopaminergic input from the substantia nigra pars compacta (SNc). The reward component of the incentive memory, barely sketched here, comprises the basolateral amygdala (BLA) and accumbens core (NACc). Reinforcement is mediated by affective modulation of SNc by the amygdala central nucleus (CeN). Further details can be found in Refs. 15, 40, 41, and 43.
of the reinforcement signal and, having a low value, will also ensure that action evaluation exerts very little if any excitatory impact on motor output.

Only the very general details of the reward system are provided here. For a more detailed discussion of the reward memory in this model, see Balleine. One final point: it should be clear that outcome-specific transfer effects could be through the S-O connections in associative memory and activation of incentive and reward processes to ensure an increased activation of the motor response. Because these effects are resistant to outcome devaluation, however, it is likely that these effects are more usually produced by stimulus-induced retrieval of the O-R association in S-R memory, resulting in an increased probability of selection and hence performance of R.

THE ROLE OF CORICOBASAL GANGLIA NETWORKS IN INSTRUMENTAL PERFORMANCE

When integrated with the existing literature, considerable recent research in our lab has allowed us to provide a reasonable working model of the neural network within which the associative–cybernetic model presented in Figure 4A could be implemented in the brain. This is presented in summary form in Figure 4B. Although we cannot describe this implementation in any detail here, its salient features and the evidence on which it is based can be quickly sketched; the interested reader is referred to the primary sources referenced later.

First a caveat: as should be clear from the above discussion, it is only possible to evaluate the effects of neural manipulations on instrumental performance when those manipulations are combined with tests that directly establish that the effects of the manipulation are specifically on instrumental conditioning and not due to changes in other forms of learning, notably Pavlovian conditioning, or due solely to changes in motor performance. Although a variety of theories have been advanced regarding the neural bases of the goal-directed aspects of instrumental conditioning, many of these have been derived using tasks that are only nominally instrumental and for which no direct evidence was provided to indicate that instrumental learning, rather than some other form of learning, was controlling performance (e.g., Refs. 44–47). Furthermore, many investigators have relied on only indirect measures of learning, for example, changes in performance or in the correlation between apparently goal-directed behavior and markers of neural activity, rather than on tests that provide direct assessments of the influence of a neural manipulation on learning.  For these reasons, much of the evidence behind these current theories is of little value in evaluating the putative mechanisms of instrumental performance.

The following summary of our working hypothesis regarding the neural bases of instrumental performance has been derived, however, from the results
of experiments designed to assess the effects of neural manipulations on specific markers of instrumental learning, that is, on the instrumental outcome devaluation and contingency degradation tests. Using these tests, we have found strong evidence suggesting that the prelimbic region of the prefrontal cortex is involved in encoding the R-O contingency. \cite{13,31,51} Cell body lesions of this structure made prior to training did not prevent the acquisition of an instrumental action but did abolish the ability of animals to choose a course of action after the outcome had been devalued or when a specific action–outcome contingency had been degraded. It is clear, however, that this region does not store the action–outcome association; after a period of training, lesions of this area no longer induced a deficit and performance appeared to be normal. \cite{31}

Furthermore, in line with the division of R-O and O-R associations, neither outcome-specific Pavlovian instrumental transfer \cite{51} nor the specificity of instrumental reinstatement was affected by prelimbic lesions, although, similar to the influence of outcome devaluation, the latter was generally attenuated by this treatment. \cite{51}

The time-limited influence of the prelimbic area on the acquisition of instrumental actions, coupled with its connectivity—particularly its amygdala, thalamic, and midbrain dopaminergic afferents—suggested that this region acts in a coordinated fashion with other prefrontal areas, notably anterior cingulate and medial agranular areas, that send efferents to common regions of thalamus and striatum to encode and modulate the encoding of the R-O association. In line with this suggestion, we found that a similar pattern of deficits was induced by lesions of the mediodorsal thalamus, a structure that maintains reciprocal connections with the prelimbic area, but not by lesions of the anterior thalamic nuclei. \cite{52} The prelimbic prefrontal cortex also sends well-documented projections to the striatum, notably the dorsomedial striatum and the core of the nucleus accumbens. Considerable evidence suggests that the latter region is not involved in instrumental learning \cite{23,53,54} although, as noted earlier, it does appear to play a role in establishing action values, likely through its connections with the basolateral amygdala, an area that we have found to be critical for establishing the incentive value of the instrumental outcome. \cite{55,56}

In contrast, we have found evidence that lesions of the posterior dorsomedial striatum, whether made prior to or after training, abolished sensitivity to both outcome devaluation and contingency degradation. \cite{57} Furthermore, temporary inactivation using muscimol or pretreatment with the N-Methyl-D-Aspartate (NMDA) antagonist 2-amino-5-phosphonopentanoic acid (APV) prior to learning were also found to abolish sensitivity to outcome devaluation but not infusions into adjacent dorsolateral striatum. \cite{57,58} Indeed, in contrast to the previously dominant view that the dorsal striatum mediates the acquisition of skills or habits, \cite{59,60} this evidence suggests that it is heterogeneous in function subserving both goal-directed and habitual actions. We have investigated this heterogeneity in several ways. First, on the basis of the observation that most studies assessing the role of the dorsal striatum in sensorimotor functions
confined their manipulations of dorsolateral striatum, we examined the effects of lesions of this area on overtraining-induced insensitivity to outcome devaluation. In this study we found that although the instrumental performance of sham lesioned rats was insensitive to outcome devaluation, lesions of this structure rendered performance once again goal directed; that is, lesioned rats showed greater control by the R-O association than sham rats. In contrast, lesions of the dorsomedial striatum were without effect. Furthermore, in a second series of experiments we found evidence that, although overtained rats were relatively insensitive to changes in the R-O contingency, muscimol-induced inactivation of the dorsolateral striatum increased the rats’ sensitivity to this manipulation.

Finally, it is important to note that the distinct reward and reinforcement functions of the outcome representation in the associative and S-R memory systems in the associative–cybernetic model overlay the distinct functions in associative and motor striatum. As mentioned earlier, the basolateral amygdala appears to play a direct role in establishing the reward value of the instrumental outcome by attaching emotional significance to its distinct sensory features. From the model, the reward memory establishes the reward value of the outcome and acts to gate the output of the associative memory on the motor system. This could be achieved either directly, through the direct connections that the basolateral area maintains with the dorsomedial striatum, the prelilbic area and mediodorsal thalamus, or indirectly through basolateral control of accumbens output onto the ventral tegmental area, the source of the dopaminergic input to dorsomedial striatum. Likewise, connections between the amygdala and accumbens shell exert control over Pavlovian–instrumental transfer (Corbit et al.) likely through the shell’s control of midbrain dopaminergic efferents on the dorsal striatum. In contrast, the role of the instrumental outcome as a reinforcer in the S-R memory is likely mediated by amygdala central nucleus efferents on the substantia nigra pars compacta, the source of the dopaminergic input to dorsolateral striatum. Current evidence suggests that this input is modulated by affective processes in the central and extended amygdala, and, as such, the amygdala as a whole appears in a position to parse the distinct reward and reinforcement signals produced by outcome delivery and to separate their impact onto the R-O and S-R associations in the parallel circuits that course through the dorsal striatum.

Generally, therefore, although the dorsomedial region of the striatum mediates the encoding of R-O associations, the dorsolateral region appears to be critical for S-R learning. In terms of the associative–cybernetic model, therefore, current evidence suggests that the critical associative and S-R memory systems that contribute to instrumental performance course through corticosstriatal circuits localized to adjacent regions of the dorsal striatum. How these two pathways interact, however, is currently a matter of debate. The generally accepted architecture of the basal ganglia emphasizes the operation of functionally distinct, closed parallel loops connecting prefrontal cortex, dorsal
striatum, pallidum/substantial nigra, thalamus, and feeding back onto the originating area of prefrontal cortex. 71–73 There is, on this view, vertical integration within loops but not lateral integration across loops and, as a consequence, various theories have had to be developed to account for lateral integration, for example, the split loop 68 or spiraling midbrain–striatal integration 35, 74 models. These models have not yet found wide acceptance. In contrast, older theories of striato-pallido-nigral integration proposed that, rather than being discrete, corticostriatal connections converge onto common target regions particularly in the globus pallidus and substantia nigra, a view that allows naturally for integration between various corticostriatal circuits. 75–77 Although anatomical studies challenge this view, recent evidence has emerged supporting a hybrid version: that, in addition to the segregated loops, there may also be integration through collateral projections from caudate (or dorsomedial striatum) converging with projections from the putamen (or dorsolateral striatum) onto common regions in both the internal and external globus pallidus. 78 Whether these converging projections underlie the integration of the O-R and R-O associations that the associative–cybernetic model identifies as critical for the initiation of instrumental performance remains an open question.

ACKNOWLEDGMENTS

The research reported here and the preparation of this manuscript were supported by a grant (no. MH56446) from the National Institute of Mental Health. The authors thank Sanne de Wit for permission to present the data in Figure 1C and Andrew Delamater for his comments on an earlier version of this manuscript.

REFERENCES


