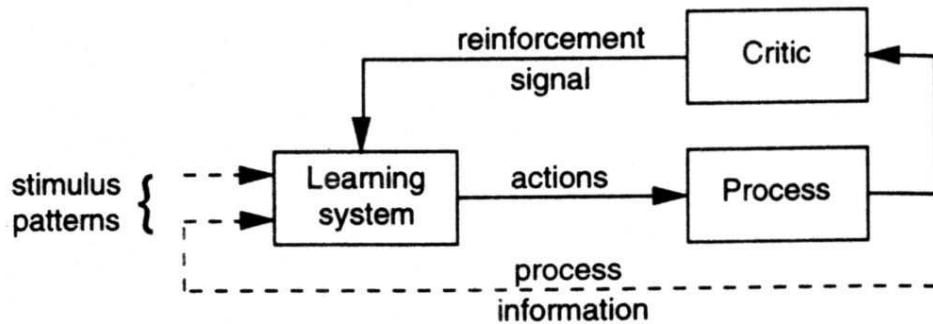
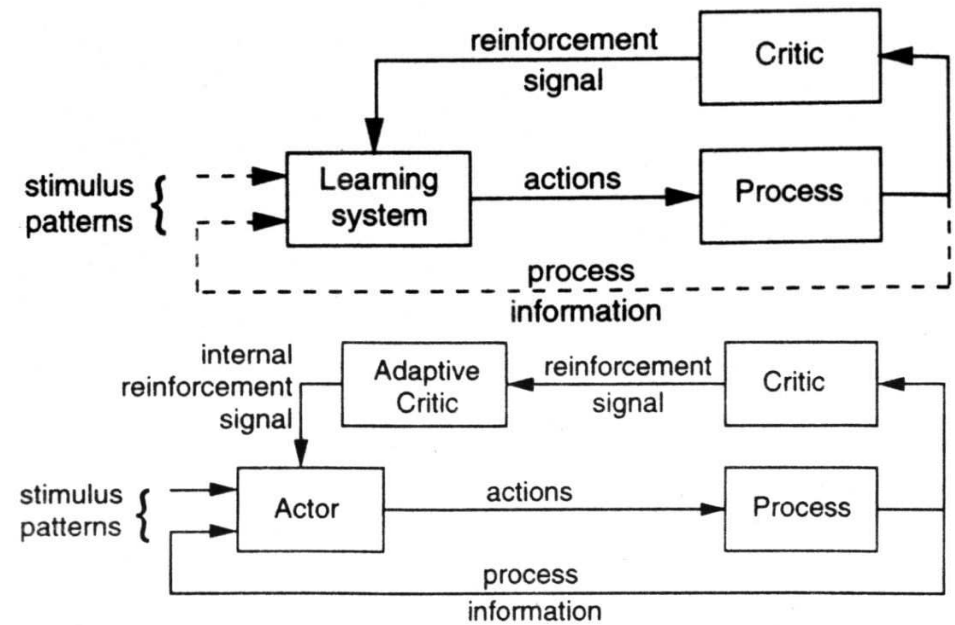


Reinforcement learning

- Optimal/effective actions are not provided to learner; must be *discovered*
- Feedback (reinforcement signal) reflects overall consequences of action (and other things) in environment
- Feedback can be intermittent, probabilistic, temporally delayed, and dependent on things outside learner's control
- Tension between *exploration* and *exploitation*



Adaptive critic



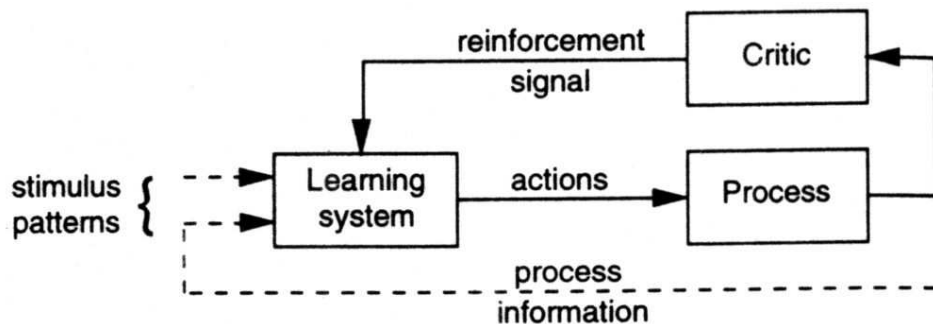
Associative reinforcement learning

- Given input, learn to produce output (action) that maximized immediate reward
- Modified Associative reward-penalty (A_{R-P})

$$p(a_j = 1) = 1 / (1 + \exp(-n_j))$$

$$\Delta w_{ij} = \rho r (a_j - p_j) a_i + \lambda \rho (1 - r) ((1 - a_j) - p_j) a_i$$

– Reinforcement is *broadcast* within multilayer network



Sequential reinforcement learning

Execute sequence of actions that maximizes *expected discounted sum* of future rewards

$$E \{ r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \dots \} = E \left\{ \sum_{k=0}^{\infty} \gamma^k r(t+k) \right\}$$

Temporal difference (TD) methods

- Learn to predict expected discounted reward

$$a_j(t+1) = E \{ r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \dots \}$$

$$a_j(t) = E \{ r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \dots \}$$

$$= E \{ r(t) \} + \gamma a_j(t+1)$$

$$\Delta w_{ij}(t) = \rho (r(t) - E \{ r(t) \}) a_i$$

$$= \rho (r(t) - (a_j(t) - \gamma a_j(t+1))) a_i$$

- Use as internal reinforcement for learning actions

Q-learning

- Learn to predict expected discounted reward associated with each action in each state, then choose best action

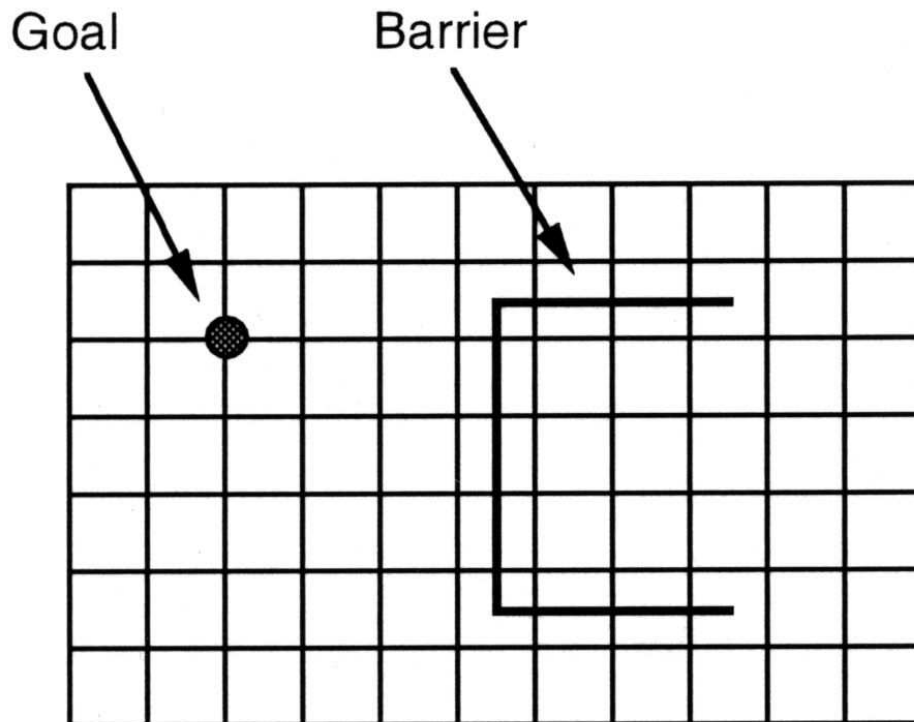
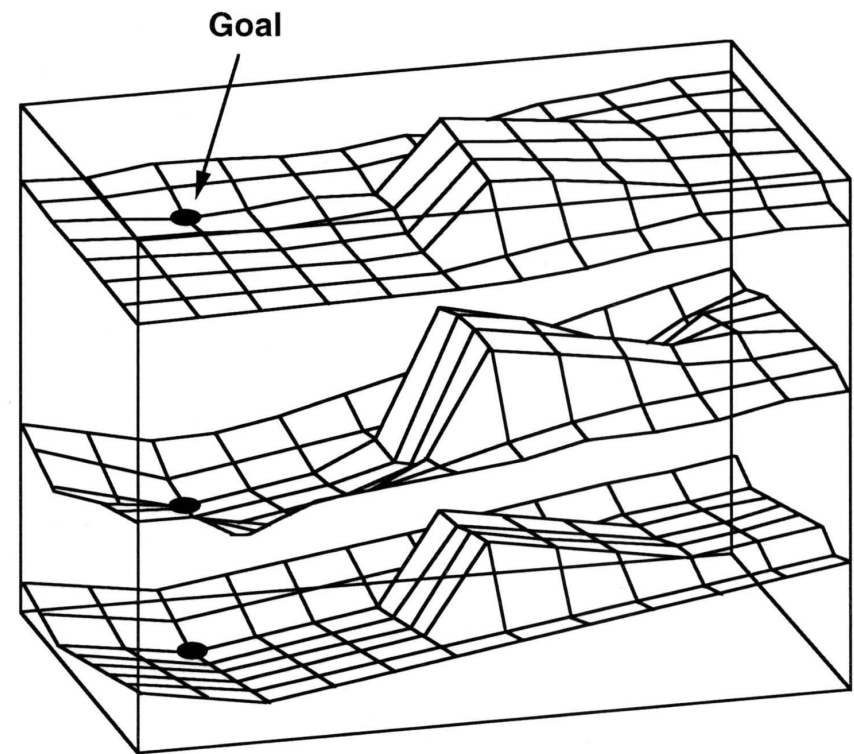
Strengths and weaknesses of reinforcement learning

Strengths

- No need for explicit behavioral targets
- Can be applied to networks of binary stochastic units
- TD can learn at least some types of temporal behavior (discrete sequential)
- Error (in associative learning) is broadcast rather than back-propagated
- TD learning consistent with some physiological evidence (Schultz et al.)
- Can use associative reinforcement learning (e.g., A_{R-P}) to learn actions based on prediction of reinforcement learned by TD.

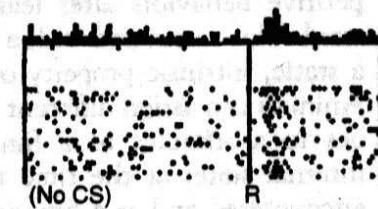
Weaknesses

- Learning is often *very* slow
- Application to large/continuous state spaces requires some mechanism for function approximation—e.g., multilayer network trained with back-propagation
- No one has successfully combined associative and TD learning (last point above) in anything but the simplest domains.

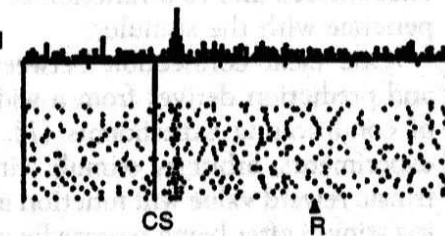


Do dopamine neurons report an error in the prediction of reward?

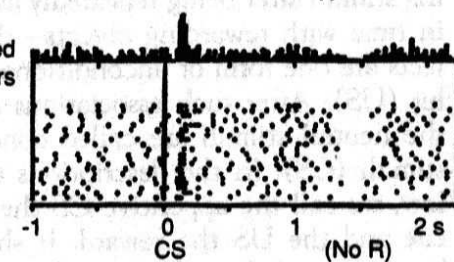
No prediction
Reward occurs



Reward predicted
Reward occurs



Reward predicted
No reward occurs



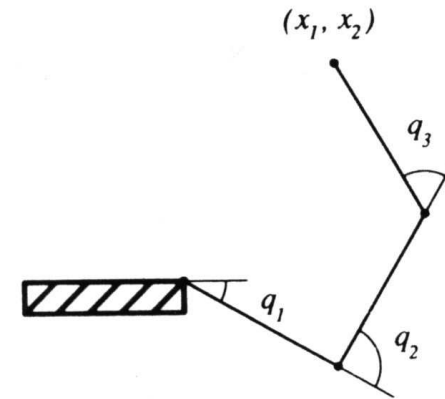
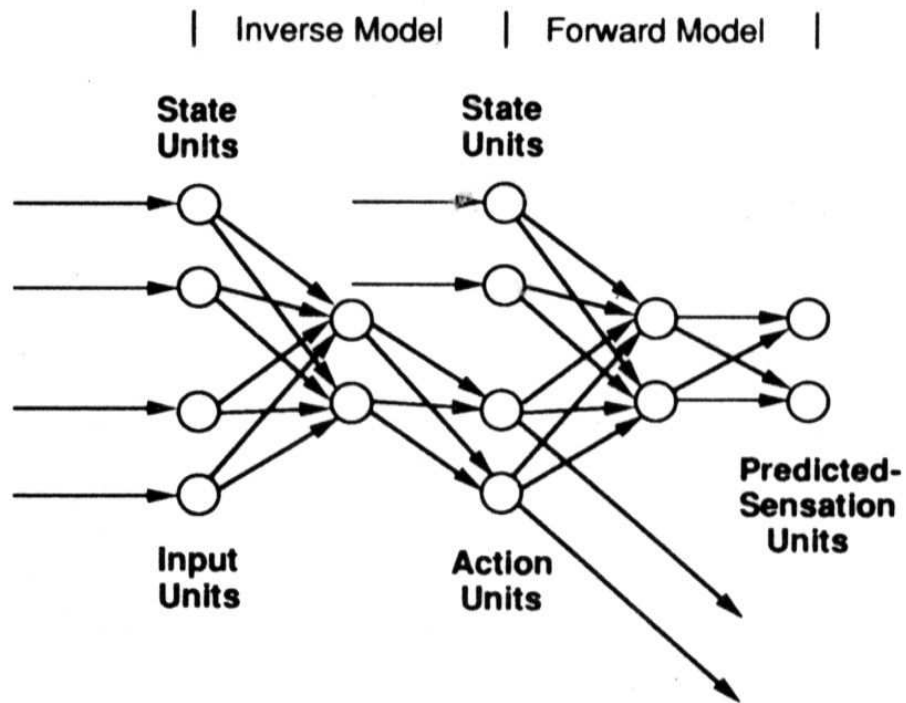
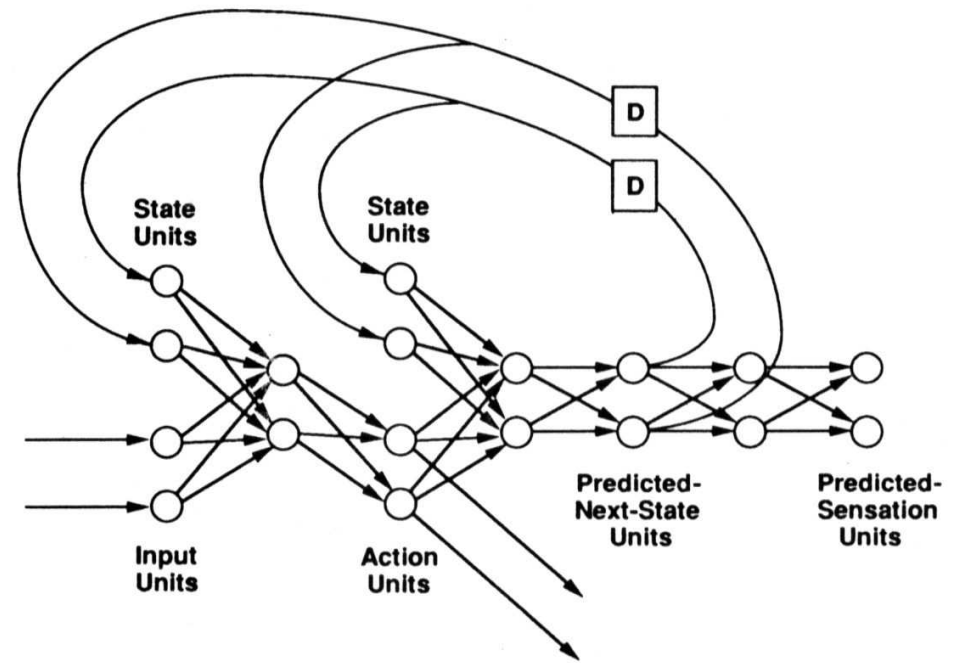
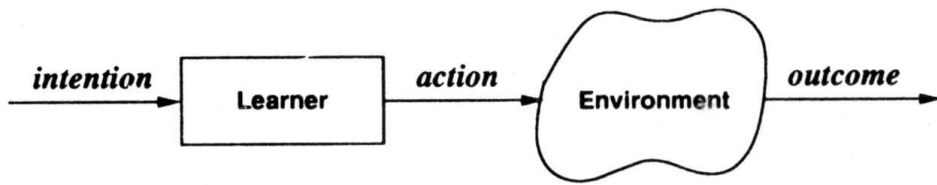
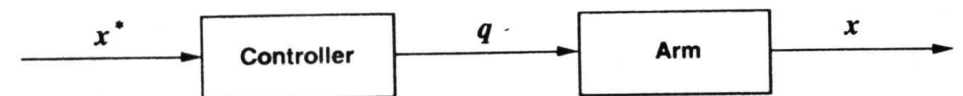
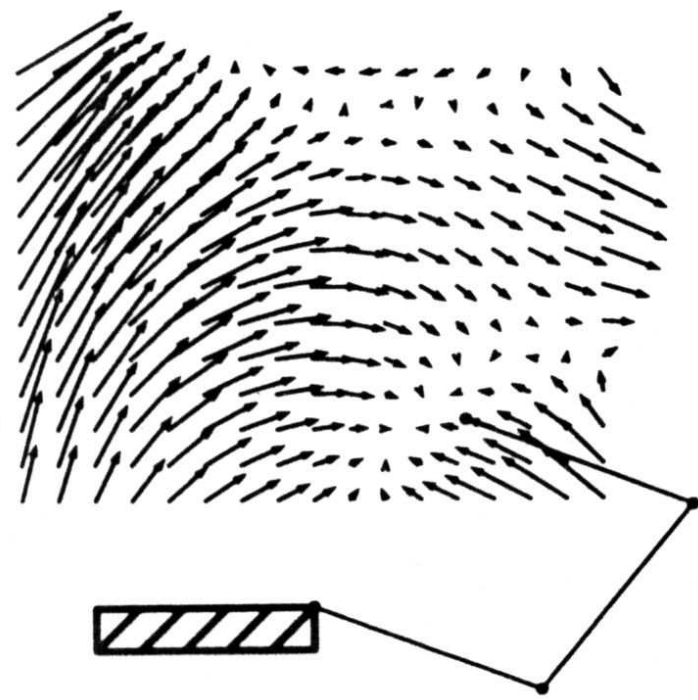
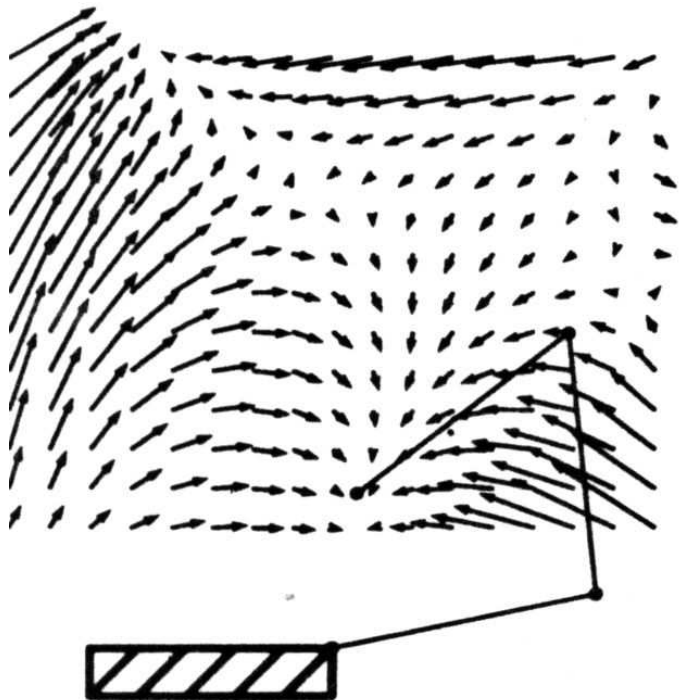
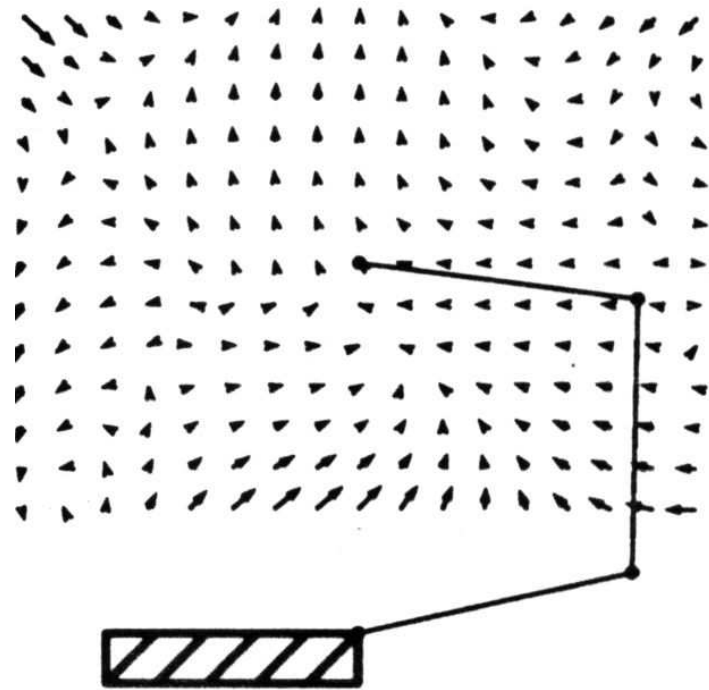
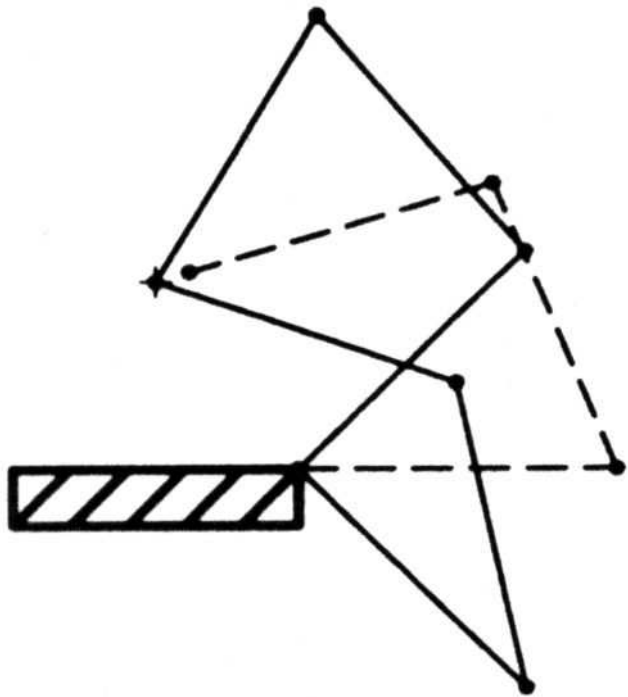


Figure 11. A three-joint planar arm.





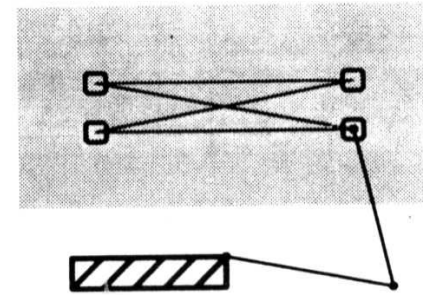
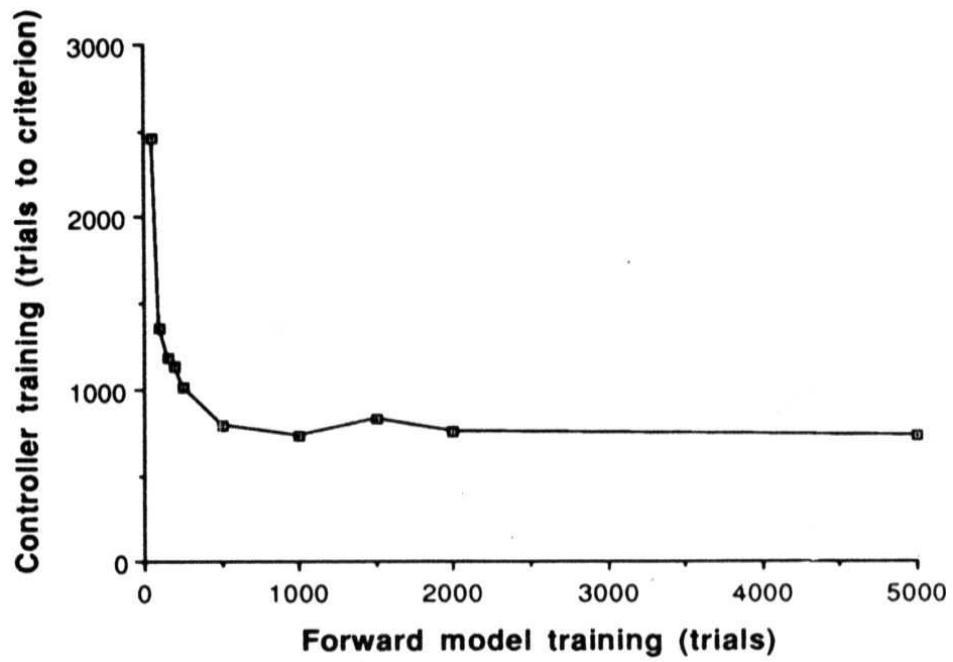


Figure 20. The workspace (the gray region) and four target paths: The trajectories move from left to right along the paths shown.

