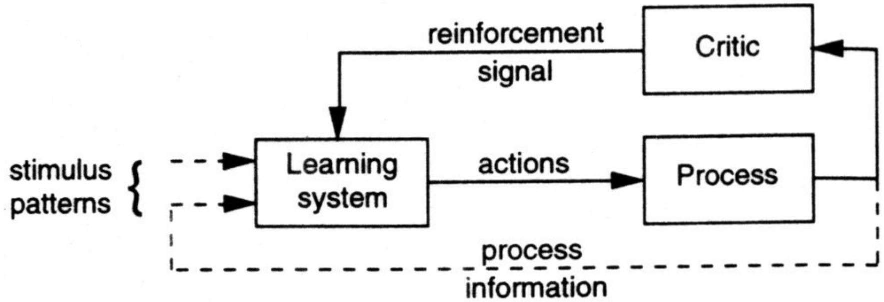
Reinforcement learning



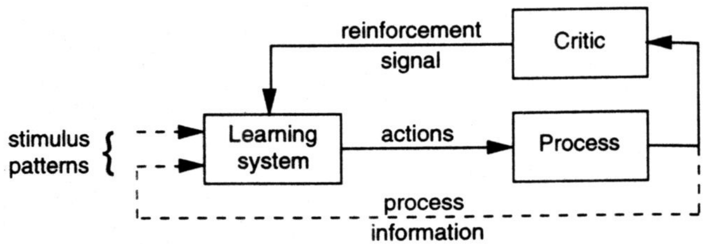
- Optimal/effective actions are not provided to learner; must be *discovered*
- Feedback (reinforcement signal) reflects overall consequences of action (and other things) in environment
- Feedback can be intermittent, probabalistic, temporally delayed, and dependent on things outside learner's control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

Associative reinforcement learning

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

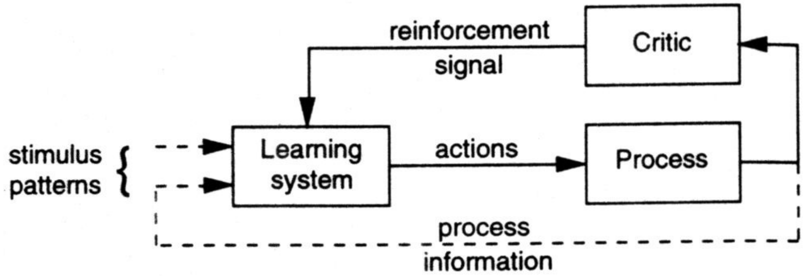
$$
\begin{aligned}
p\,(a_j = 1) &= 1/\left(1 + \exp(-n_j)\right) \\
\Delta w_{ij} &= \begin{cases} \rho\,(\quad a_j \quad - n_j)\, a_i & \text{if } \textit{success} \\ \lambda\rho\,((1 - a_j) - n_j)\, a_i & \text{if } \textit{failure} \end{cases}
\end{aligned}
$$

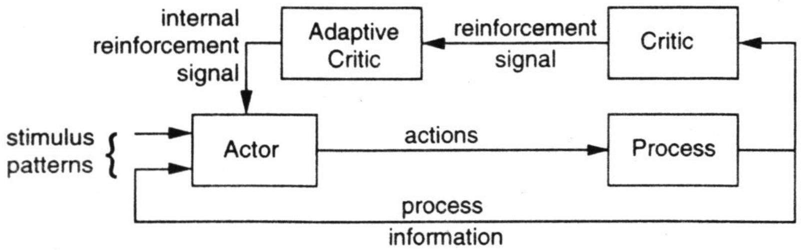
  – Reinforcement is *broadcast* within multilayer network



Adaptive critic

- Feedback can be intermittent, probabalistic, temporally delayed





Sequential reinforcement learning

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

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

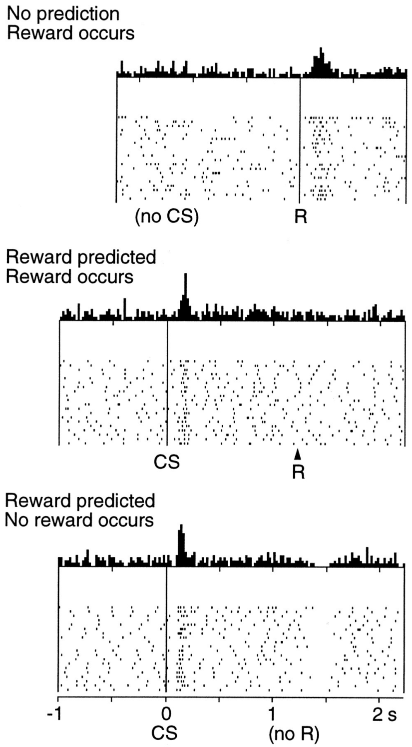
- Temporal difference (TD) methods
  – Learn to predict expected discounted reward

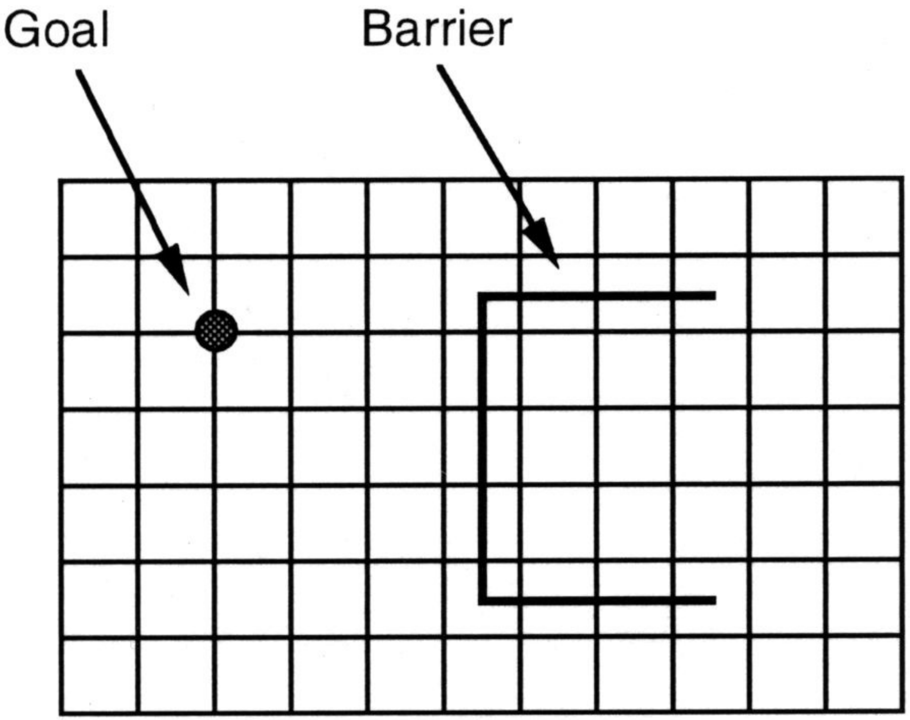
$$
\begin{aligned}
a_j(t+1) &= E\left\{\qquad r(t+1) + \gamma\; r(t+2) + \gamma^2 r(t+3) + \cdots\right\} \\
a_j(t) &= E\left\{r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) \cdots\right\} \\
&= E\left\{r(t)\right\} + \gamma a_j(t+1) \\
E\left\{r(t)\right\} &= a_j(t) - \gamma a_j(t+1) \\
\\
\Delta w_{ij}(t) &= \rho\,(r\,(t) - \quad E\left\{r\,(t)\right\} \quad)\, a_i \\
&= \rho\,(r\,(t) - (a_j\,(t) - \gamma a_j\,(t+1))\ )\, a_i
\end{aligned}
$$

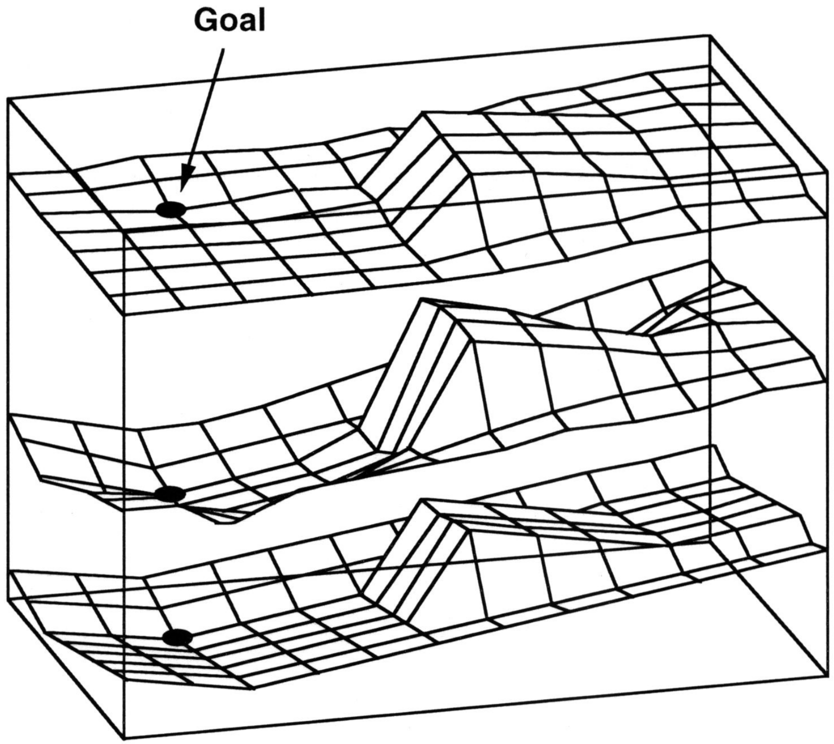
  – Use as internal reinforcement for learning actions

Dopamine and reward prediction (Shultz et al., 1997)

- Classical conditioning
- Response of dopaminergic neurons in **substantia nigra** (subcortical nucleus)







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
- Associative learning error is broadcast rather than back-propagated
- TD learning consistent with some physiological evidence (Schultz)
- Can use associative reinforcement learning (e.g., $A_{R\text{-}P}$) to learn actions based on prediction of reinforcement learned by TD

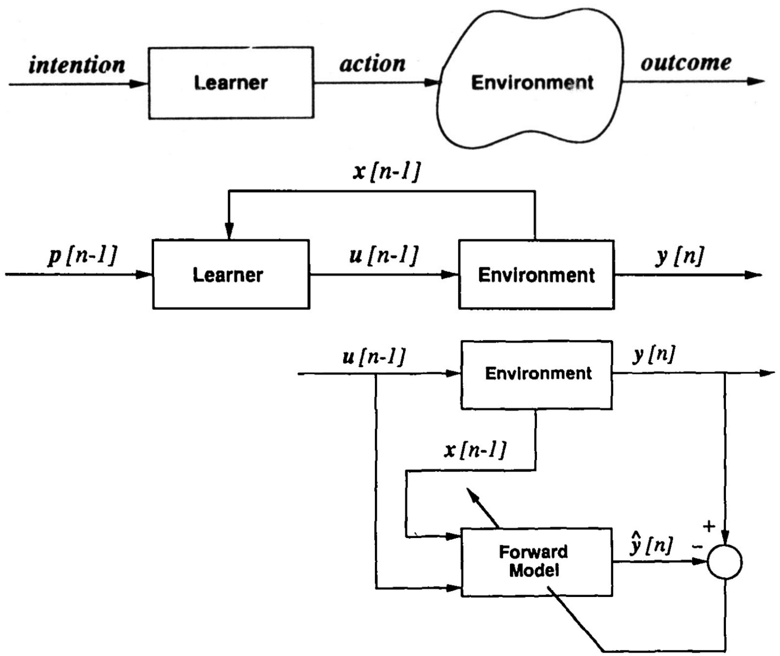
Weaknesses

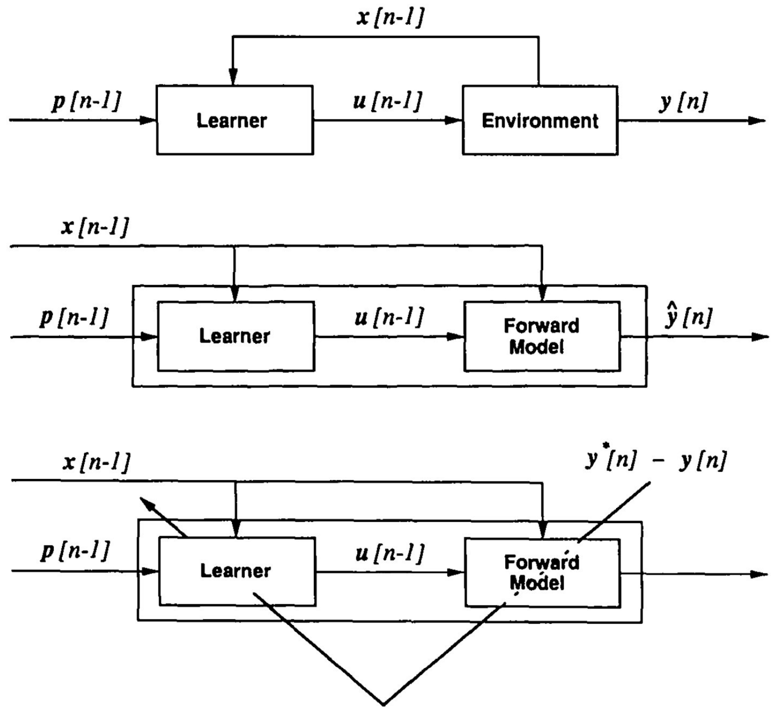
- Learning is often *very* slow
- Application to large/continuous state spaces requires some mechanism for function approximation—e.g., multilayer network trained with back-propagation
- Associative and TD learning combined only in very simple domains

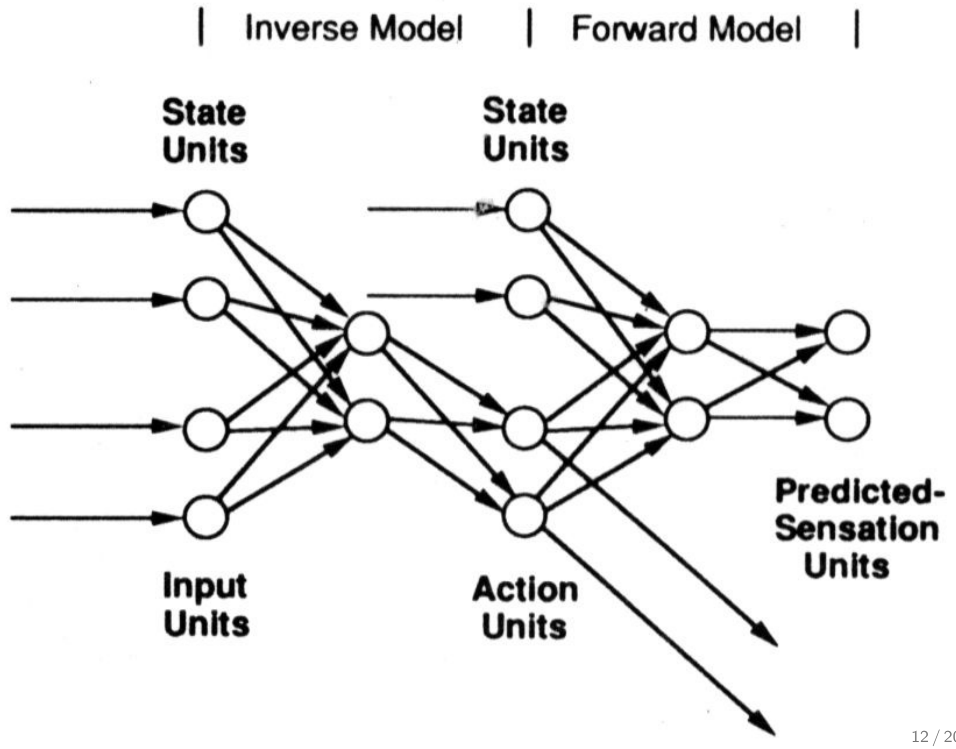
Forward models

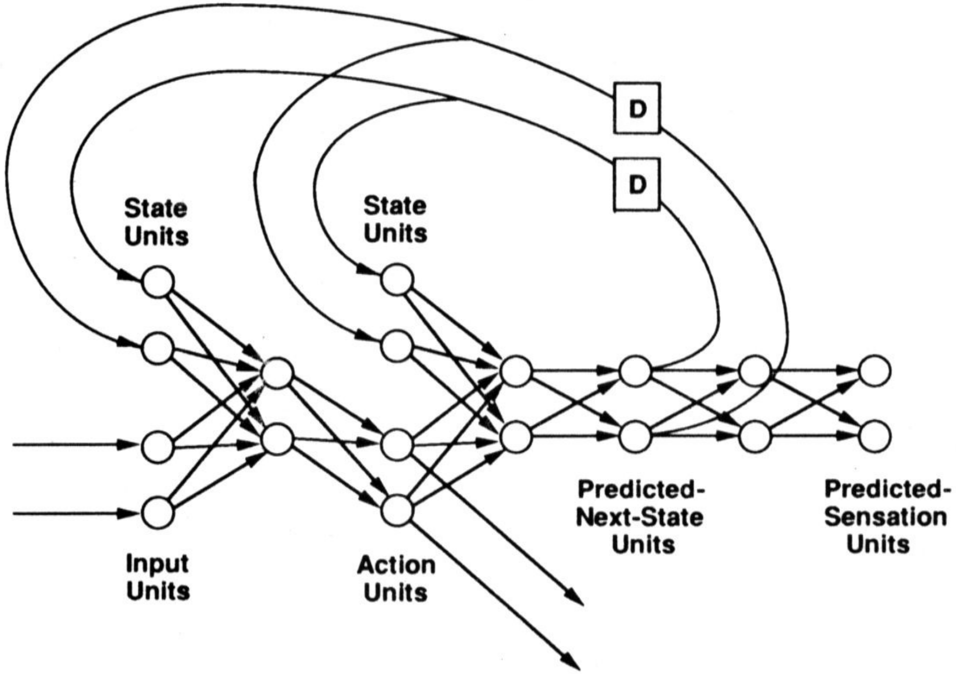
- Feedback from the world is in terms of *distal* error (observable consequences) rather than *proximal* error (motor commands)
- Would like compute proximal error from distal error (to improve motor commands to achieve goals)
- Relationship between motor commands and observable consequences involves processes in the external world (e.g., physics)
- Learn an internal (forward) model of the world which can be *inverted* (e.g., back-propagated through) to convert distal error to proximal error
    - Such a model can also provide online outcome prediction to detect errors during execution

Forward models









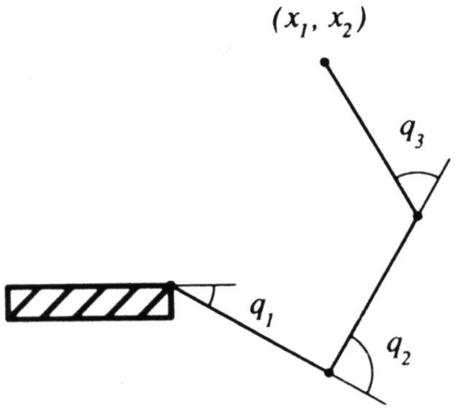
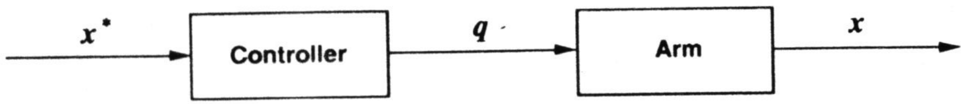
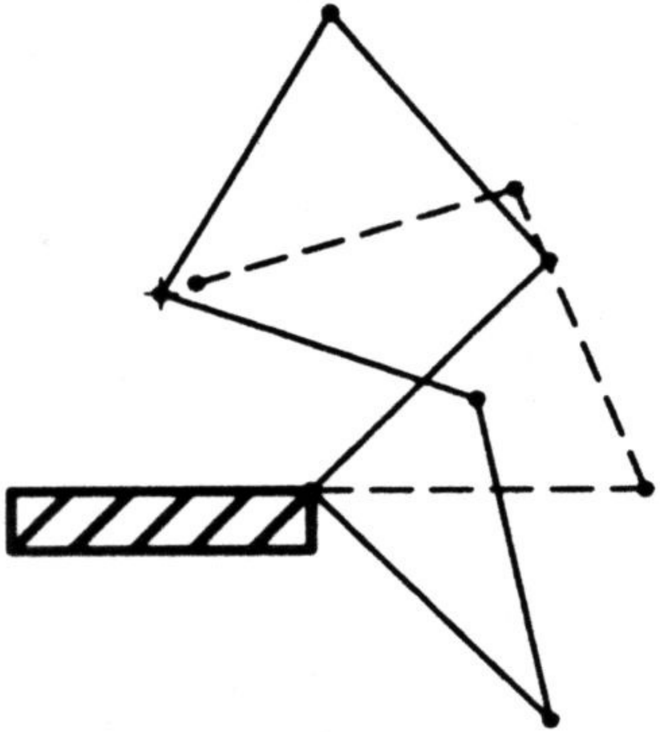
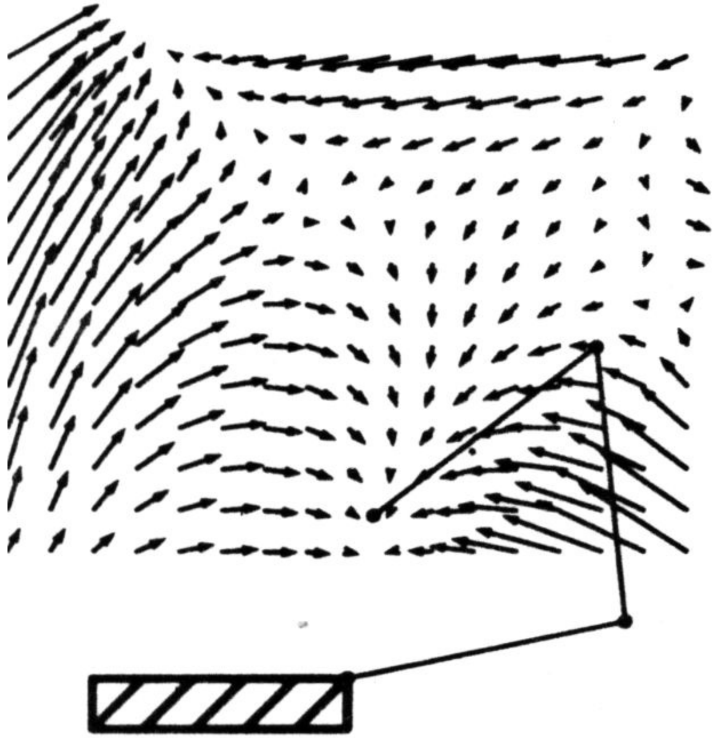
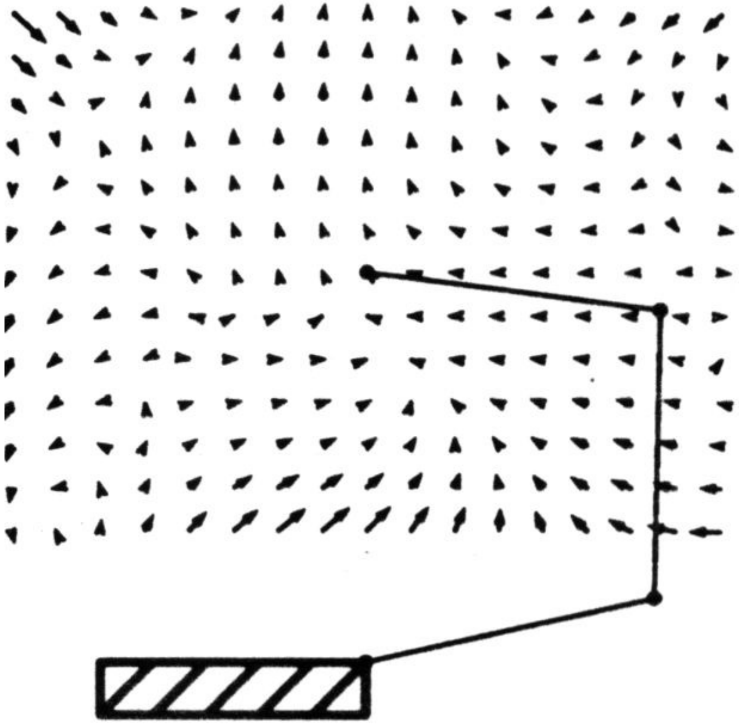


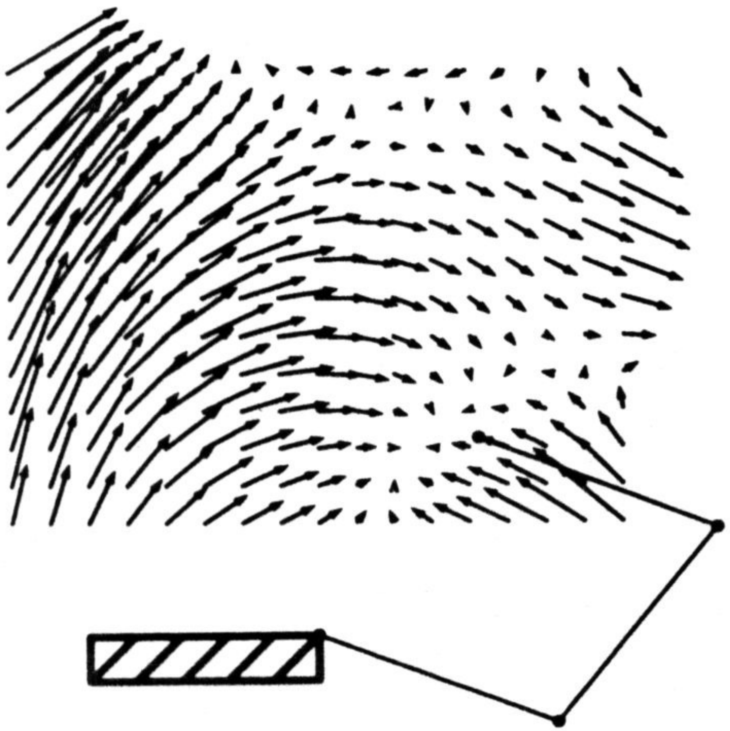
Figure 11. A three-joint planar arm.

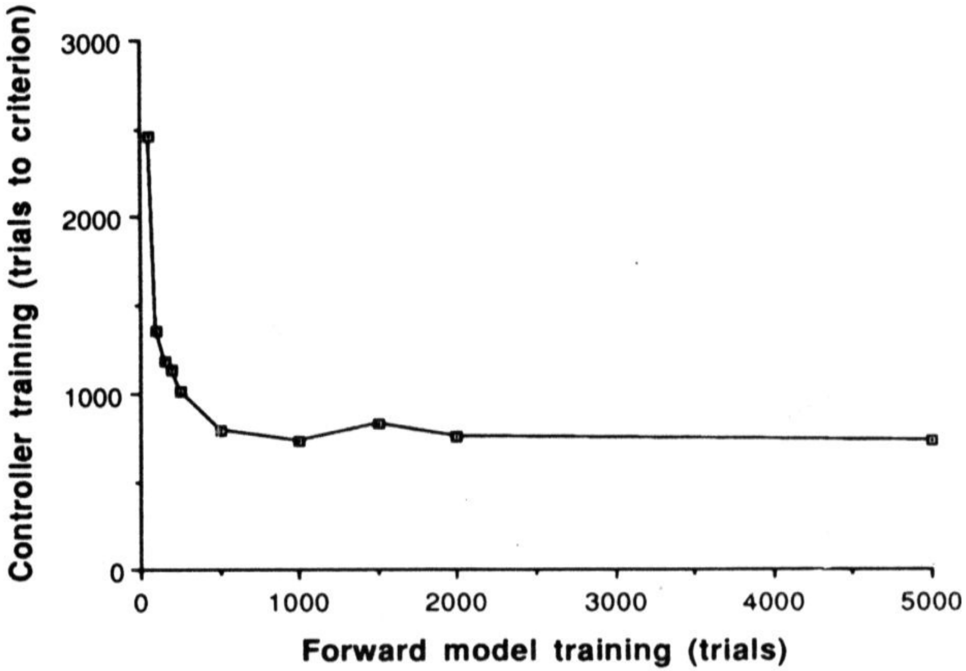












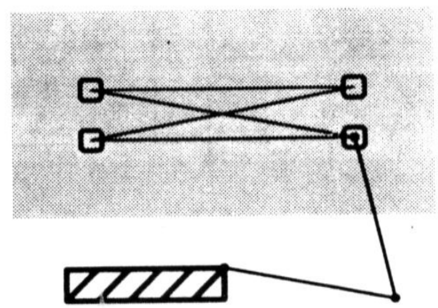


Figure 20. The workspace (the gray region) and four target paths: The trajectories move from left to right along the paths shown.