

# Reinforcement Learning Training 2025

# Bellman's Equation

Foundation to solving MDP and RL problems.

## Recall (1)

- Markov decision process has transition probabilities

$$\Pr \left[ S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a \right]$$

which transitions the agent to state  $S_{t+1}$  and a reward of  $R_{t+1}$

- Cumulative reward at time  $t$

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

## Recall (2)

- A value function is an expected cumulative reward

$$v_{\pi}(s) = \mathbb{E}_{\pi}[G_t | S_t = s]$$

- An action-value function is an expected cumulative reward from taking action  $a$

$$q_{\pi}(s, a) = \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a]$$

■ Note that  $v$  and  $q$  depend on the policy  $\pi$ .

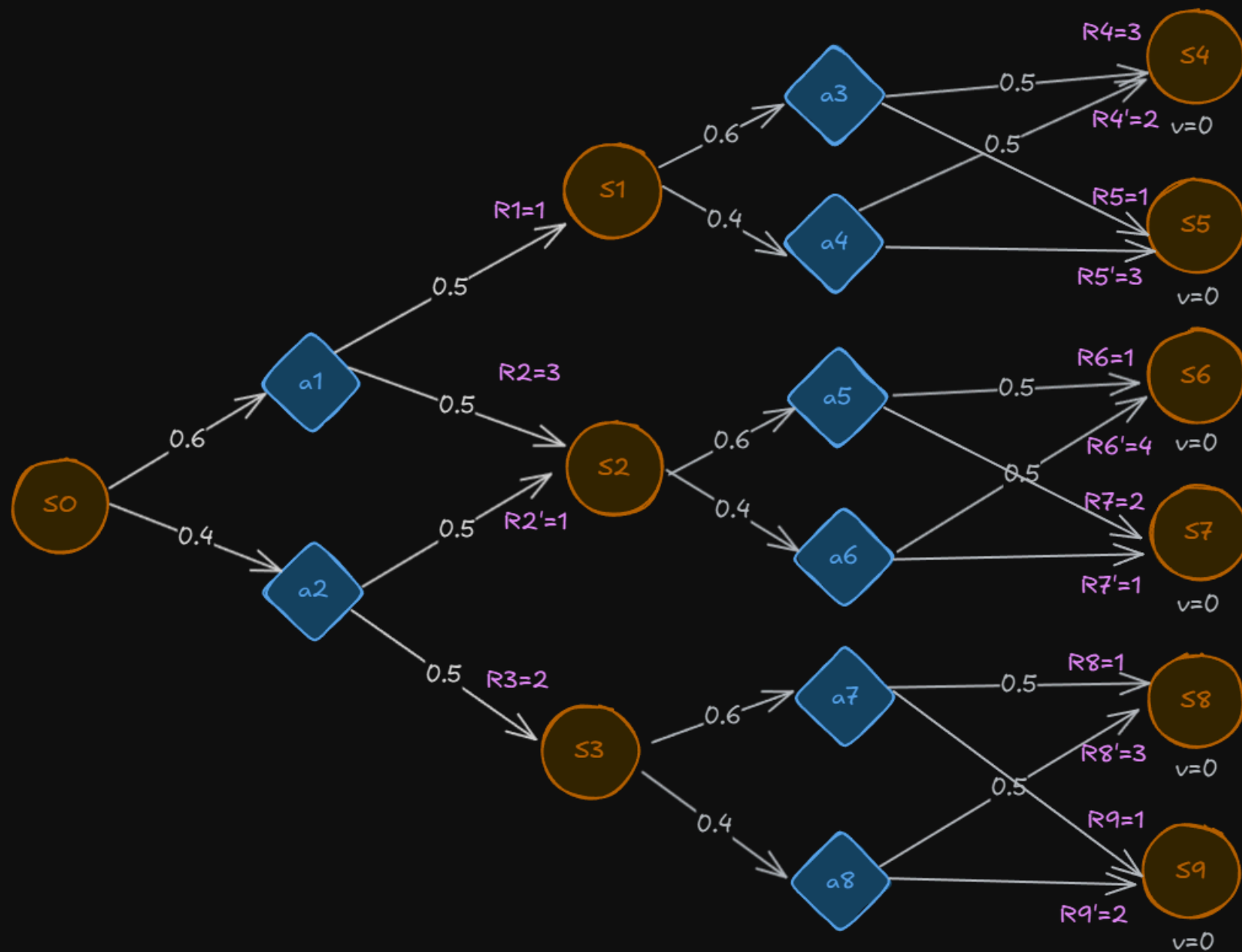
# Bellman Equation

- Allows relationships among  $v$  and  $q$ .

$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma v_{\pi}(s')]$$

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) \left[ r + \gamma \sum_{a'} \pi(a'|s') q(s', a') \right]$$

# Example



**Find**  $v(s_1)$

$$V(s_1) = 0.6 \left[ 0.5 \times (R_4 + \gamma \cancel{V(s_4)}) + 0.5 \times (R_5 + \gamma \cancel{V(s_5)}) \right] \\ + 0.4 \left[ 0.5 \times (R'_4 + \gamma \cancel{V(s_4)}) + 0.5 \times (R'_5 + \gamma \cancel{V(s_5)}) \right]$$

$$V(s_1) = 0.6 \left[ 0.5(3) + 0.5(1) \right] + 0.4 \left[ 0.5(2) + 0.5(3) \right] \\ = 2.2$$

**Find  $v(s_2)$  and  $v(s_3)$**

$$\begin{aligned}V(s_2) &= 0.6 [0.5(1) + 0.5(2)] + 0.4 [0.5(4) + 0.5(1)] = 1.9 \\V(s_3) &= 0.6 [0.5(1) + 0.5(1)] + 0.4 [0.5(3) + 0.5(2)] = 1.6\end{aligned}$$

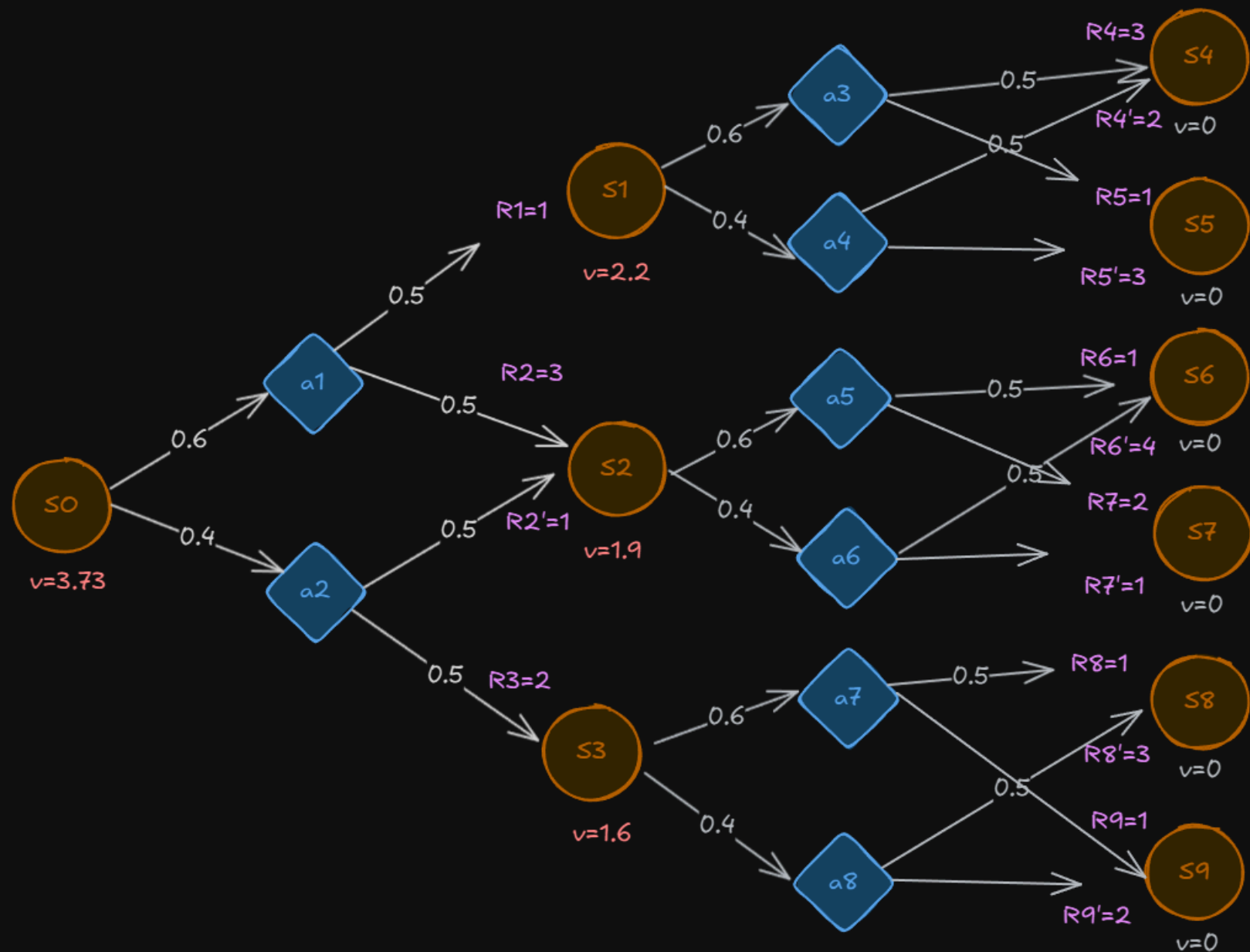


**Find**  $v(s_0)$

$$\begin{aligned} V(s_0) &= 0.6 \left[ 0.5(1 + 2.2) + 0.5(3 + 1.9) \right] \\ &\quad + 0.4 \left[ 0.5(1 + 1.9) + 0.5(2 + 1.6) \right] \\ &= 0.6(4.25) + 0.4(3.25) \\ &= 3.93 \end{aligned}$$

# Result

See values of  $v$



**Find  $q(a_3)$**

$$\begin{aligned} q(a_3) &= 0.5 \left[ R_4 + \gamma \sum_{a'} \cancel{(\dots)} \right] + 0.5 \left[ R_5 + \gamma \sum_{a'} \cancel{(\dots)} \right] \\ &= 0.5(3) + (0.5)(1) \\ &= 2 \end{aligned}$$

*Handwritten notes: "0 (No action)" with an arrow pointing to the first summation, and "0" with an arrow pointing to the second summation.*

**Find**  $q(a_4) - q(a_8)$

$$q(a_4) = 0.5(2) + 0.5(3) = 2.5$$

$$q(a_5) = 0.5(1) + 0.5(2) = 1.5$$

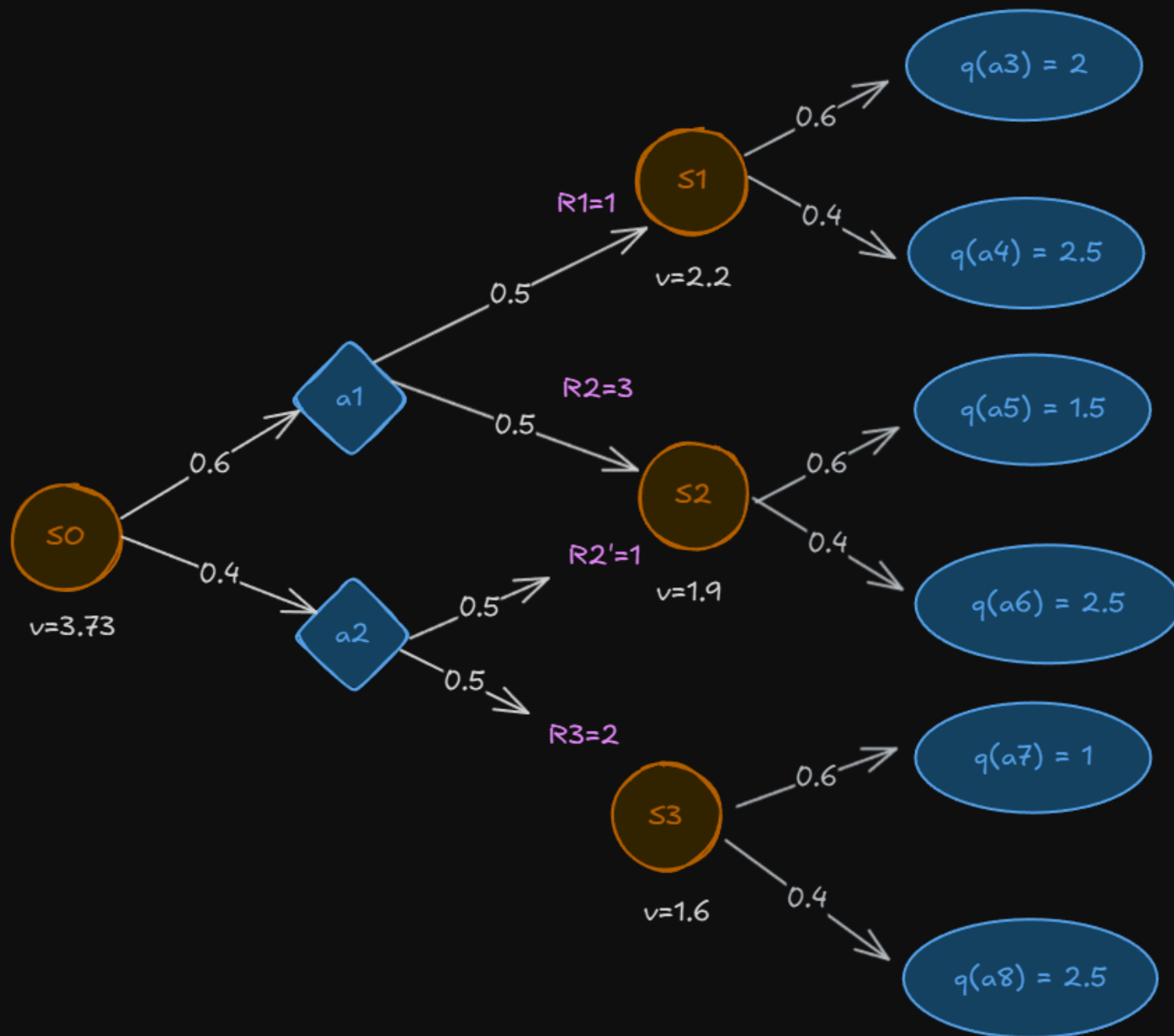
$$q(a_6) = 0.5(4) + 0.5(1) = 2.5$$

$$q(a_7) = 0.5(1) + 0.5(1) = 1$$

$$q(a_8) = 0.5(3) + 0.5(2) = 2.5$$

# Result

See  $q$

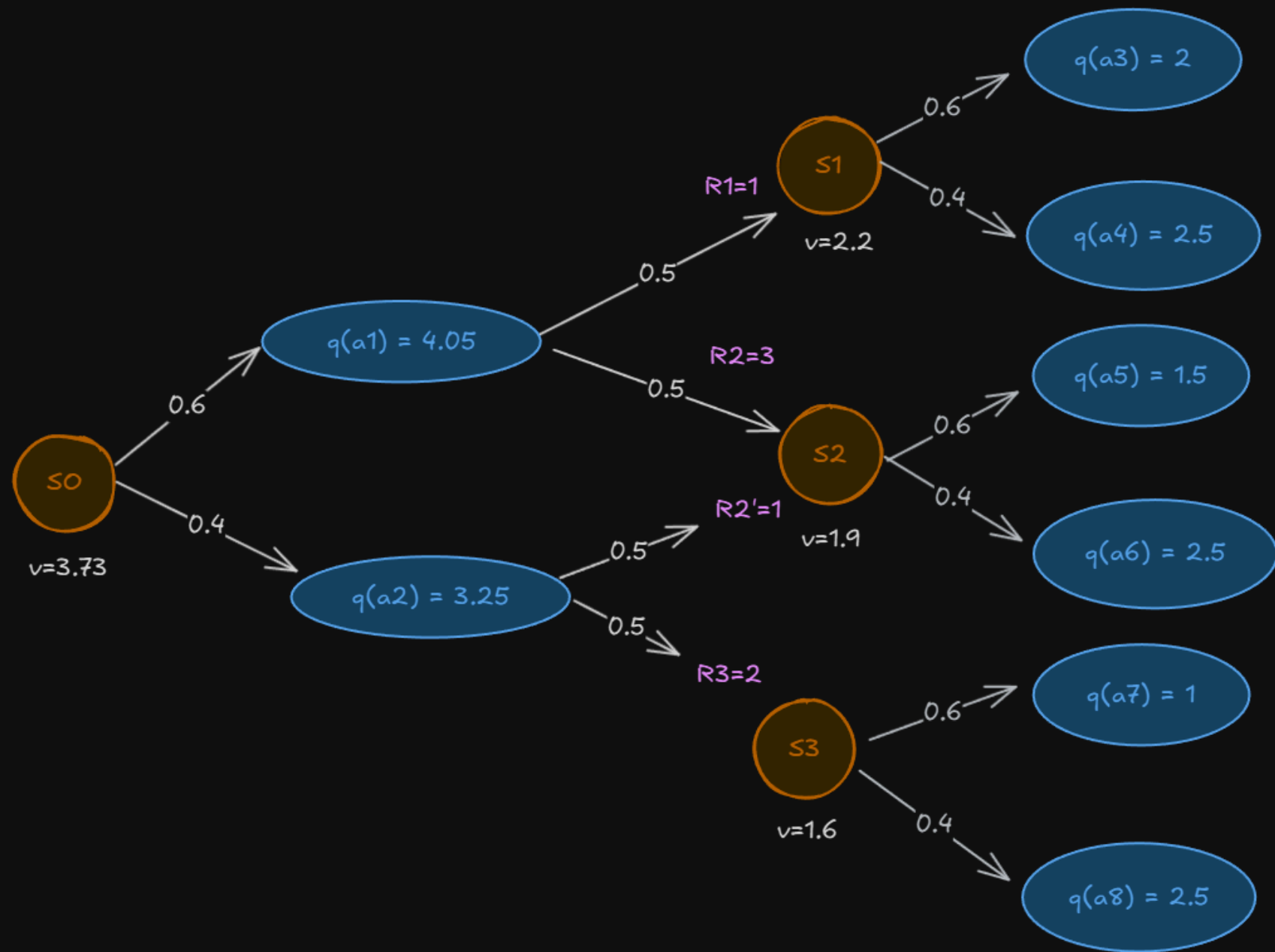


**\$ Find  $q(a_1), q(a_2)$**

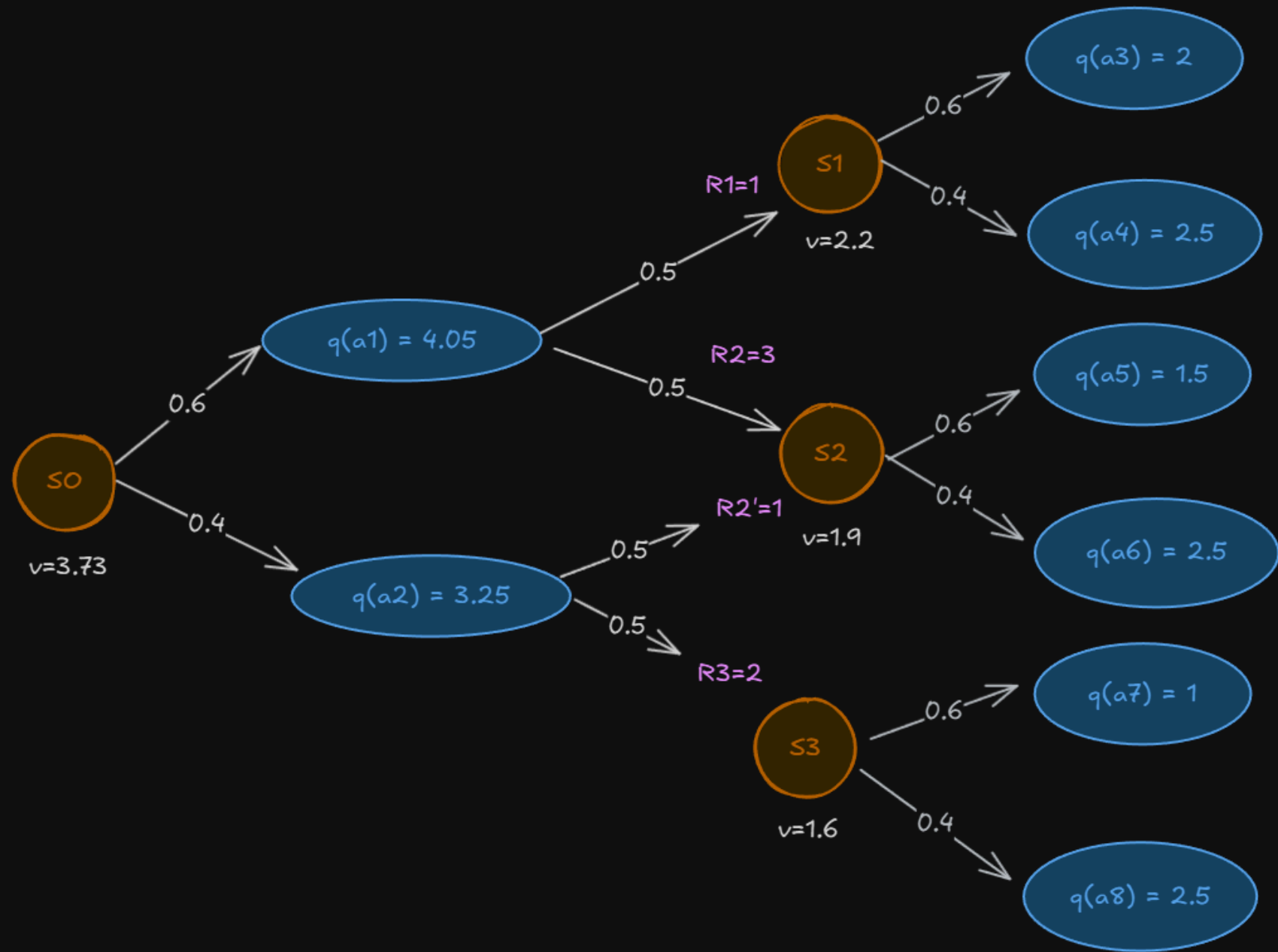
$$\begin{aligned} q(a_1) &= 0.5 \left[ 1 + \cancel{\gamma} (0.6(2) + 0.4(2.5)) \right] \\ &\quad + 0.5 \left[ 3 + \cancel{\gamma} (0.6(1.5) + 0.4(2.5)) \right] \\ &= 4.05 \end{aligned}$$

$$\begin{aligned} q(a_2) &= 0.5 \left[ 1 + \gamma (0.6(1.5) + 0.4(2.5)) \right] \\ &\quad + 0.5 \left[ 2 + \gamma (0.6(1) + 0.4(2.5)) \right] \\ &= 3.25 \end{aligned}$$

# Result

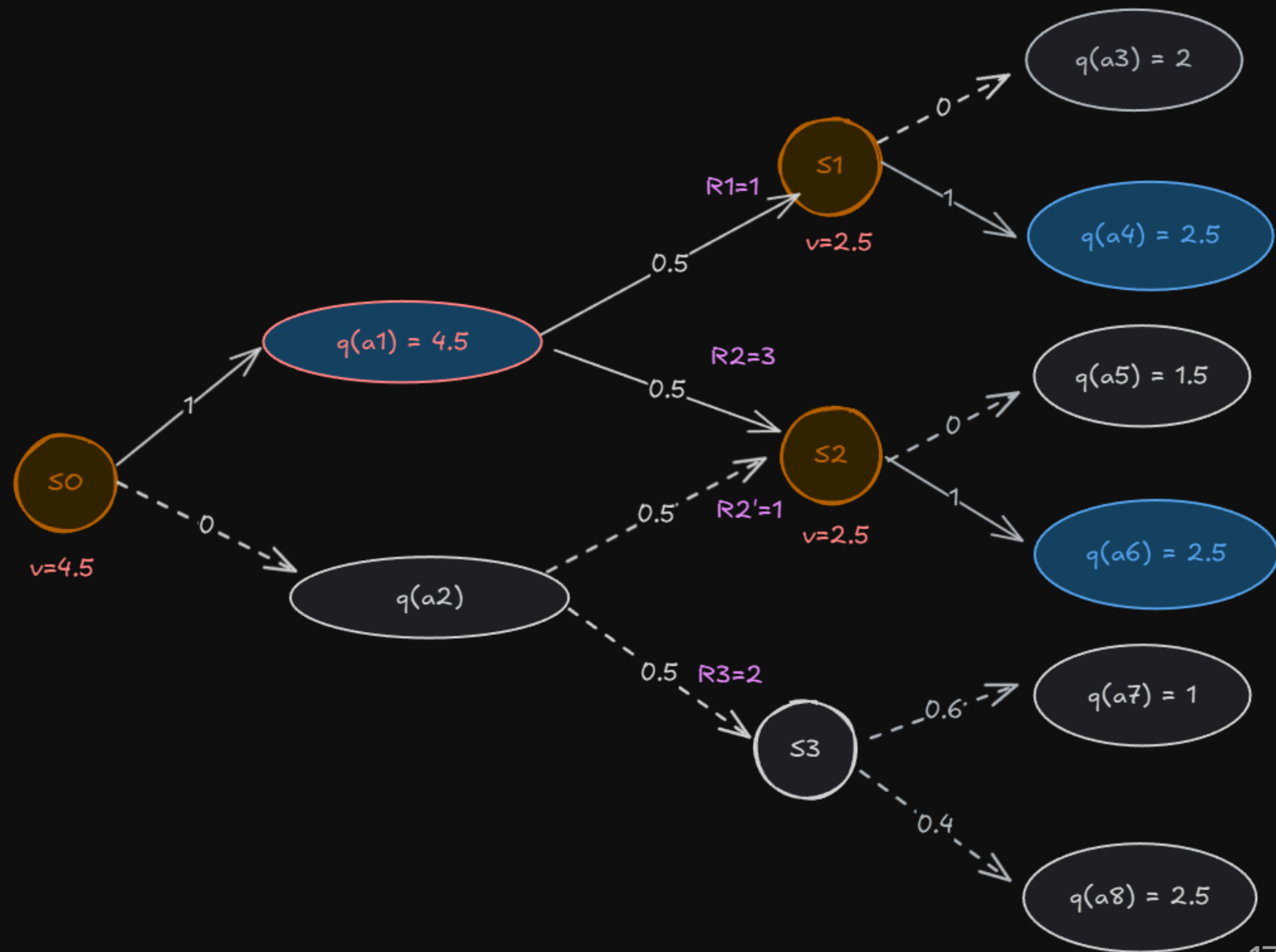


Can we do better?





# Optimal Policy



# Optimality Condition

- If agent is following the optimal policy  $\pi^*$  (something we want to find), then the value function will also be optimal.

$$v^*(s) = \max_{\pi} v_{\pi}(s)$$

- It follows that

$$v^*(s) = \max_a \left\{ \sum_{s', r} p(s', r | s, a) [r + \gamma v^*(s')] \right\}$$

$$q^*(s, a) = \sum_{s', r} p(s', r | s, a) \cdot \left[ r + \gamma \max_{a'} q^*(s', a') \right]$$