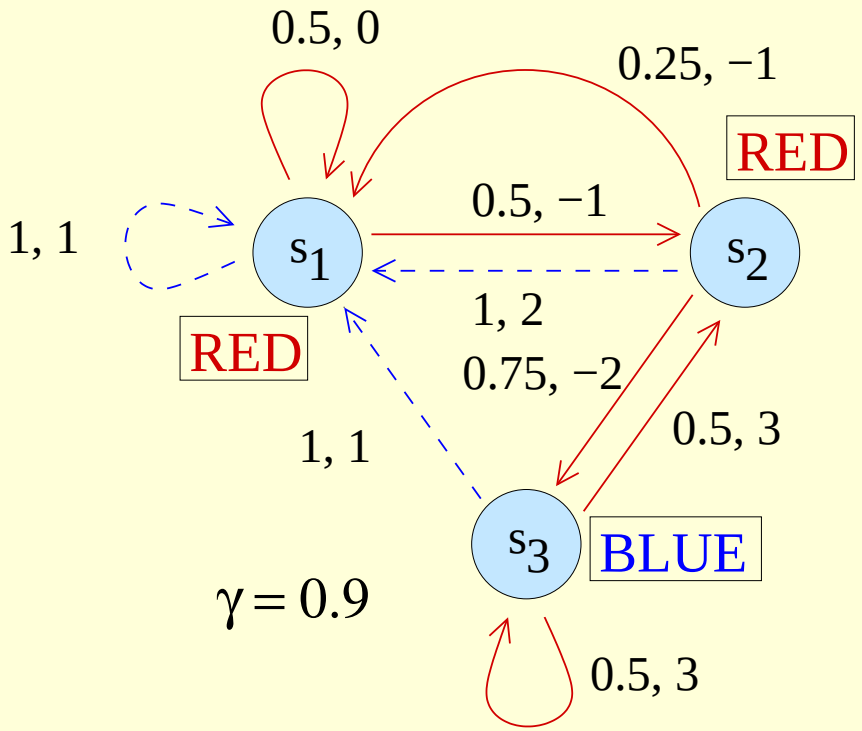
Markov Decision Problems (MDPs)



Elements of an MDP

- States (S)
- Actions (A)
- Transition probabilities (T)
- Rewards (R)
- Discount factor (γ)

Behaviour is encoded as a **Policy** π, which maps states to actions.
What is a "good" policy? One that maximises expected long-term reward.

V^π is the **Value Function** of π. For $s \in S$,

$$V^\pi(s) = \mathbb{E}_\pi \left[r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots | \text{start state} = s \right].$$

Optimal Policies

V^π satisfies a recursive equation: $V^\pi = R_\pi + \gamma T_\pi V^\pi$, which gives $V^\pi = (I - \gamma T_\pi)^{-1} R_\pi$.

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$	
RRR	4.45	6.55	10.82	
RRB	-5.61	-5.75	-4.05	
RBR	2.76	4.48	9.12	
RBB	2.76	4.48	3.48	
BRR	10.0	9.34	13.10	
BRB	10.0	7.25	10.0	
BBR	**10.0**	**11.0**	**14.45**	← Optimal policy
BBB	10.0	11.0	10.0	

Every MDP is guaranteed to have an optimal policy $\pi^\star$, such that

$$\forall \pi \in \Pi, \forall s \in S : V^{\pi^\star}(s) \geq V^\pi(s).$$