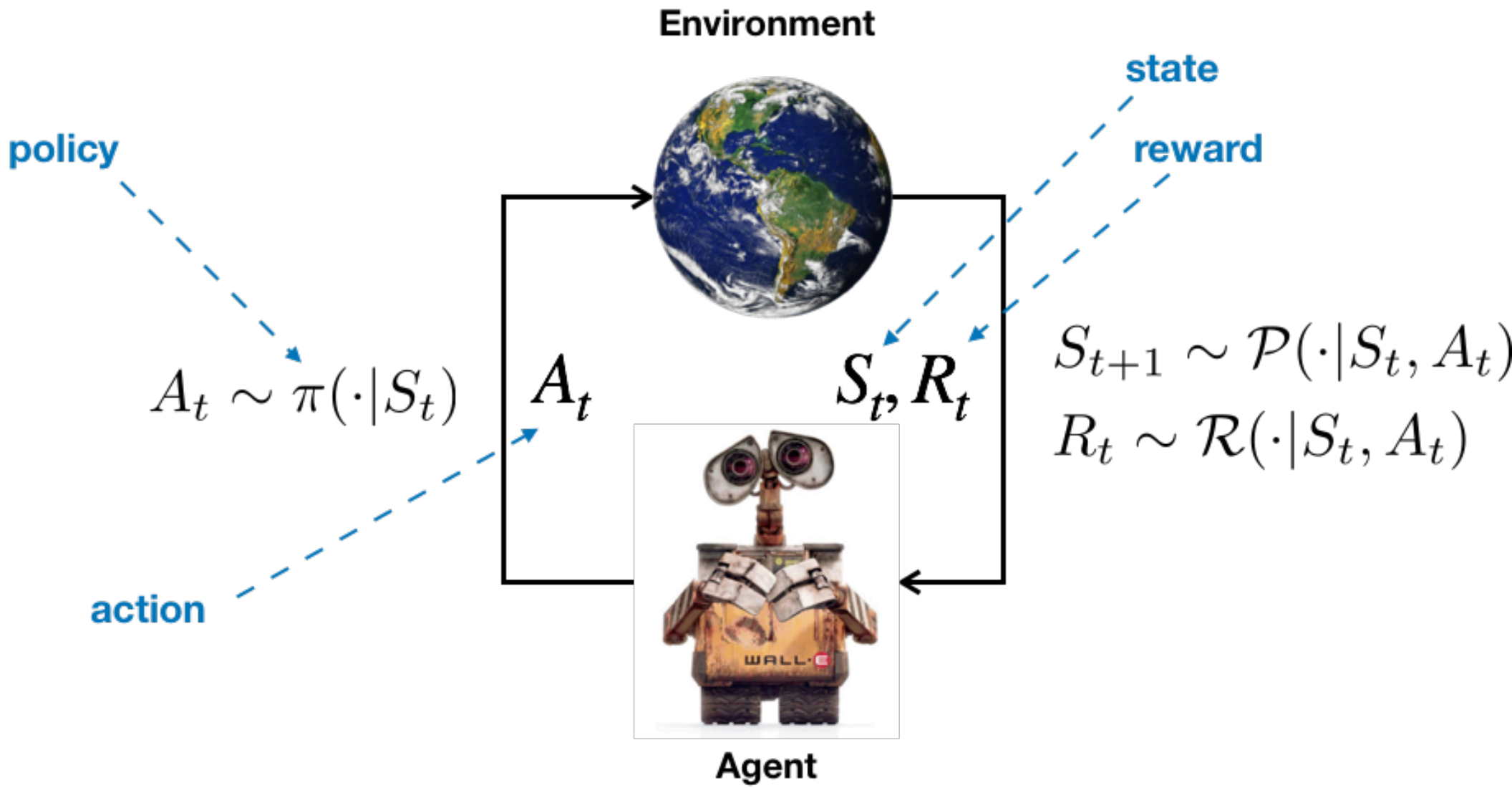
Environment
state
reward
policy
$A_t \sim \pi(\cdot|S_t)$
A_t
S_t, R_t
$S_{t+1} \sim \mathcal{P}(\cdot|S_t, A_t)$
$R_t \sim \mathcal{R}(\cdot|S_t, A_t)$
action
Agent