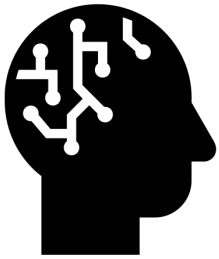
方策(p)に応じて行動(a)を選択

エージェント



環境

累積報酬を最大化
する方策を学習

行動に応じた
- 次の状態(s)
- 報酬(r)