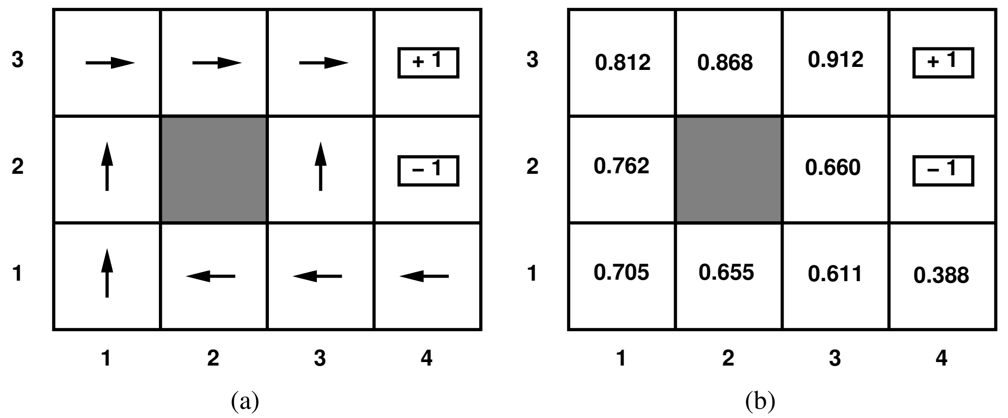

Figure 17.2 (a) An optimal policy for the stochastic environment. (b) The utilities of the states.

function SIMPLE-POLICY-AGENT(*percept*) **returns** an *action*
  static: *M*, a transition model
    U, a utility function on environment histories
    P, a policy, initially unknown

  if *P* is unknown **then** *P* ← the optimal policy given *U*, *M*
  return *P*[*percept*]

Figure 17.3 An agent that calculates and uses an optimal policy.

is equipped with a sonar ring that gives it the distance to the nearest wall in each of the four directions. For such an agent, the locations (2,1) and (2,3) are indistinguishable, yet different actions are needed in each. Furthermore, the Markov property does not hold for percepts (as opposed to states), because the next percept does not depend just on the current *percept* and the action taken.

The correct approach for POMDPs is to calculate a probability distribution over the possible states given all previous percepts, and to base decisions on this distribution. Although the optimal decision is not uniquely determined by the current percept, it *is* determined uniquely (up to ties) by the agent's probability distribution over the possible states that it could be in. For example, the sonar-equipped agent might believe that it is in state (2,1) with probability 0.8 and in state (2,3) with probability 0.2. The utility of action A is then

$0.8 \times$ utility of doing A in (2,1) given current beliefs +
$0.2 \times$ utility of doing A in (2,3) given current beliefs

This seems simple enough. Unfortunately, in POMDPs, calculating the utility of an action in a state is made more difficult by the fact that actions will cause the agent to obtain new percepts, which will cause the agent's beliefs to change in complex ways that depend on its current beliefs.