

Kvíz na určení policy a ohodnocení stavů

-60	-1	80
-5	-1	50

Čísla označují rewards (odměny). Červeně označené stavy jsou terminální.

Agent se může pohybovat všemi 4 směry. Agent poslechne s pravděpodobností 0.6 a s pravděpodobností 0.4 se vydá opačným směrem. Koefficient zapomínání (discount factor) je pro jednoduchost výpočtů 1, okamžitá odměna (immediate reward) je pro každý stav uvedena v tabulce. Pro neterminální stavy spočítejte hodnoty (values/utilities) $V(s)$ s přesností ± 1 (max. odchylka od skutečné hodnoty) a optimální policy.