Kvíz na určení policy a ohodnocení stavů

50	$R_B = -1$	-5
80	$R_A = -1$	-60

Stavy A a B jsou neterminální a mají reward -1. Červeně označené stavy jsou terminální; čísla zobrazují odměny.

Agent se může pohybovat všemi 4 směry. Agent poslechne s pravděpodobností 0.6 a s pravděpodobností 0.4 se vydá opačným směrem. Koeficient 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 vyznačte optimální strategii/policy.