Alex Hendrix - A02365406

MDP formulation (Tic-Tac-Toe vs random opponent)

- State space S
  - Each state is the full board encoded as a 9-character string (row-major) with characters in {X,O,~} (or equivalently a 3×3 array of the same symbols). Example state: "X~O~~X~~~". This is the key used in the implementation (get_state_str).
- Action space A(s)
  - For a nonterminal state s, the legal actions are the indices 0–8 of empty cells: $A(s) = \{i \in \{0..8\} \mid s[i] = \sim\}$. The agent (X) picks one index per step.
- Transition model P(s' | s, a)
  - Agent X places its mark at index a deterministically, producing s1 = place(s,a,X). If s1 is terminal (win/draw) the episode ends and s' = s1.
  - Otherwise the opponent O chooses uniformly from legal actions in s1 and places an O (random move), producing s' (stochastic). Thus P is deterministic for X's immediate effect and stochastic for the environment response. Formally:
    - If X's move ends the game then P(s' | s,a) = 1 for s' = s1 and 0 otherwise.
    - Else P(s' | s,a) = 1/|A(s1)| for each s' that results from one legal opponent move in s1; 0 otherwise.
- Reward R(s,a,s')
  - Terminal reward only on transitions that end the episode:
    - +1 if X wins, −1 if O wins, 0 for a draw.
  - For nonterminal transitions R(s,a,s') = 0. (This is what your step() returns.)
- Episodic setting and discount factor
  - Episodes end at terminal states (X win, O win, or draw). Discount factor $\gamma = 1$ (undiscounted episodic MDP).