Off-policy là gì?

- **Off-policy learning**: Học giá trị dựa trên **chính sách mục tiêu (target policy)** khác với **chính sách hành vi (behavior policy)**.
    - **Chính sách hành vi**: Cách agent thực tế chọn hành động (vd: ε-greedy).
    - **Chính sách mục tiêu**: Chính sách mà agent muốn học (thường là greedy).
- Ví dụ: Q-learning học theo chính sách greedy (tối ưu), nhưng hành động theo ε-greedy (để khám phá).

Tại sao Q-learning là off-policy mà không cần importance sampling?

- Vì Q-learning **không ước lượng giá trị kỳ vọng dưới chính sách hành vi**, mà **trực tiếp học giá trị tối ưu Q***.
- Công thức cập nhật:
- $$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$
- Q-learning **luôn bootstraps từ hành động tốt nhất** (greedy), nên không cần dùng hệ số hiệu chỉnh (importance sampling).