1. Động lực: Giải quyết vấn đề khám phá mà không đánh đổi hiệu suất

- Trong **exploration-exploitation trade-off**, agent phải đánh đổi giữa **thu thập phần thưởng tối đa** và **khám phá hành động mới**.

- Tuy nhiên, **Off-Policy Learning** cho phép agent học được giá trị của **chính sách tối ưu (target policy)** mà **không cần hành động theo nó → không cần đánh đổi hiệu suất** khi hành động.

---

2. Khái niệm chính:

- **Target Policy (π)**:
  → Là chính sách mà agent muốn học.
  → Ví dụ: chính sách tối ưu.

- **Behavior Policy (b)**:
  → Là chính sách mà agent **thực sự hành động theo** để sinh dữ liệu.
  → Ví dụ: chính sách chọn hành động ngẫu nhiên để đảm bảo khám phá.

---

3. Ưu điểm của Off-Policy Learning:

- **Tách biệt việc học và hành động** → Agent có thể **khám phá hiệu quả** với behavior policy mà vẫn **học được target policy** tốt.

- Ứng dụng:

    - **Khám phá liên tục.**

    - **Học từ dữ liệu có sẵn (demonstration).**

    - **Học song song nhiều chính sách.**

---

4. Yêu cầu quan trọng: Coverage Condition

- Chính sách hành vi **phải "bao phủ"** chính sách mục tiêu:

Nếu target policy π chọn hành động a tại trạng thái s với xác suất > 0
→ thì behavior policy b **cũng phải** chọn hành động đó với xác suất > 0.

- Nếu không, **agent sẽ không có dữ liệu để ước lượng giá trị đúng cho π** → dẫn đến **sai lệch nghiêm trọng** trong việc học.

---

5. So sánh với On-Policy:

On-Policy Learning là **trường hợp đặc biệt của Off-Policy**, khi:

```
π == b
```

-
- → Off-Policy Learning **tổng quát hơn**, linh hoạt hơn trong nhiều bài toán thực tế.

---

6. Kết luận:

- Off-Policy Learning là một phương pháp học giá trị **từ dữ liệu do chính sách khác sinh ra**.
- Phân biệt:
    - **Target policy (π)** – chính sách cần học.
    - **Behavior policy (b)** – chính sách dùng để lấy dữ liệu.
- Điều kiện bắt buộc: **b phải bao phủ π** để việc học chính xác.