RL là một nhánh của học máy, nơi một **agent** (tác nhân) học cách đưa ra quyết định trong môi trường (environment) để **tối đa hóa tổng phần thưởng (reward)** trong dài hạn.

### **Quá trình học:**

1. Agent quan sát trạng thái s hiện tại từ môi trường.
2. Agent chọn một hành động a theo một chính sách (policy).
3. Môi trường phản hồi bằng phần thưởng r và trạng thái mới s'.
4. Agent sử dụng thông tin này để cập nhật kiến thức của mình (value hoặc Q-value).