Fairness, bias and ethics

- ML/AI đang ảnh hưởng đến **hàng tỷ người**: từ hệ thống tuyển dụng, nhận diện khuôn mặt, đến duyệt vay ngân hàng, nội dung hiển thị trên mạng xã hội...
- Đã có nhiều **trường hợp gây hậu quả nghiêm trọng**:
  - Công cụ tuyển dụng phân biệt giới tính.
  - Nhận diện khuôn mặt nhầm lẫn người da màu với tội phạm.
  - Hệ thống duyệt vay thiên vị, gây bất công cho nhóm thiểu số.
  - Deepfake giả mạo người nổi tiếng (Obama).
  - Bot phát tán tin giả, lời nói căm thù, gian lận tài chính...

Nguyên nhân:

- ML học theo **dữ liệu đầu vào**, nên nếu dữ liệu mang thành kiến → mô hình cũng học thành kiến.
- Tối ưu hóa cho **engagement, lợi nhuận**, có thể dẫn đến lan truyền nội dung độc hại.

Một số nguyên tắc hướng dẫn xây dựng hệ thống ML công bằng & đạo đức hơn:

1. Đa dạng hóa đội ngũ (diverse team)

- Đa dạng về giới tính, sắc tộc, văn hóa, trải nghiệm sống...
- Đội nhóm đa dạng → dễ nhận diện rủi ro gây hại tới nhóm dễ bị tổn thương.
- Giúp phát hiện thiên lệch tiềm ẩn **trước khi triển khai**.

2. Tìm hiểu tiêu chuẩn đạo đức trong ngành

- Ví dụ: ngành tài chính đã bắt đầu có chuẩn đánh giá hệ thống duyệt vay là "công bằng".
- Tìm kiếm tiêu chuẩn, tài liệu học thuật hoặc thực tiễn có thể giúp bạn làm đúng hơn.

3. Kiểm toán hệ thống trước khi triển khai (audit)

- Nếu nghi ngờ hệ thống thiên vị giới tính, sắc tộc, độ tuổi... → đo lường hiệu năng chia theo từng nhóm nhỏ (subgroup).
- **Không triển khai nếu còn thiên lệch nghiêm trọng.**

4. Lập kế hoạch ứng phó (mitigation plan)

- Ví dụ: nếu nhận thấy hệ thống hoạt động sai sau khi triển khai → quay lại dùng hệ thống cũ, hoặc kích hoạt biện pháp khẩn cấp.
- Tất cả các nhóm phát triển xe tự lái đều có **kịch bản ứng phó tai nạn** trước khi chạy thử ngoài đời.

5. Giám sát liên tục sau triển khai

- Mô hình tốt hôm nay có thể sẽ **hoạt động kém ngày mai** khi dữ liệu thực tế thay đổi (ví dụ: tên mới nổi, sự kiện mới, khủng hoảng xã hội...).
- Cần theo dõi dữ liệu đầu vào, đầu ra, và chuẩn bị cập nhật mô hình khi cần.