Tên : Đào Nguyễn Nhật Anh

MSSV: 23110073

Một số yêu cầu chính về đạo đức cho AI theo Liên minh Châu Âu (EU)

Liên minh Châu Âu (EU) đã công bố Bộ Hướng dẫn Đạo đức cho Trí tuệ nhân tạo (AI) vào năm 2019, nhấn mạnh nguyên tắc 'AI đáng tin cậy' (Trustworthy AI). Theo đó, một hệ thống AI cần đảm bảo tuân thủ pháp luật, phù hợp với các nguyên tắc đạo đức, và đủ mạnh mẽ để tránh gây hại. Dưới đây là bảy yêu cầu chính cùng ví dụ minh họa:

1. Tính chủ thể và giám sát của con người (Human agency and oversight)

AI phải hỗ trợ, chứ không thay thế, quyền tự chủ của con người. Ví dụ: trong y tế, hệ thống chẩn đoán AI cần đưa ra gợi ý cho bác sĩ, nhưng bác sĩ mới là người ra quyết định cuối cùng.

2. Độ tin cậy kỹ thuật và an toàn (Technical robustness and safety)

AI cần an toàn, đáng tin cậy và có khả năng xử lý lỗi. Ví dụ: xe tự lái phải có chế độ 'dừng khẩn cấp' để con người kiểm soát khi gặp sự cố.

3. Quyền riêng tư và quản trị dữ liệu (Privacy and data governance)

AI phải tuân thủ GDPR, bảo vệ dữ liệu cá nhân. Ví dụ: một ứng dụng chăm sóc sức khỏe phải mã hóa dữ liệu bệnh nhân và chỉ dùng cho mục đích y tế.

4. Minh bạch (Transparency)

AI cần giải thích được cách thức hoạt động và quyết định. Ví dụ: hệ thống tuyển dụng dùng AI phải có khả năng giải thích lý do từ chối ứng viên.

5. Đa dạng, không phân biệt đối xử và công bằng (Diversity, non-discrimination and fairness)

AI không được thiên vị hoặc tạo ra bất công. Ví dụ: AI trong ngân hàng phải đảm bảo không phân biệt giới tính hoặc sắc tộc khi xét duyệt khoản vay.

6. Phúc lợi xã hội và môi trường (Societal and environmental well-being)

AI cần phục vụ lợi ích xã hội và bảo vệ môi trường. Ví dụ: hệ thống AI tối ưu hóa năng lượng giúp giảm khí thải nhà kính.

7. Trách nhiệm giải trình (Accountability)

Phải có cơ chế kiểm tra, đánh giá và quy trách nhiệm khi AI gây hậu quả. Ví dụ: nếu AI trong thương mại điện tử đưa ra quyết định sai gây thiệt hại, doanh nghiệp phải có trách nhiệm bồi thường và minh bạch trong xử lý sự cố.

Tóm lại, EU hướng tới phát triển AI đáng tin cậy dựa trên giá trị nhân văn, bảo đảm quyền lợi con người và xã hội. Việc tuân thủ 7 yêu cầu trên sẽ giúp AI trở thành công cụ phục vụ con người một cách an toàn, minh bạch và công bằng.