CHƯƠNG 1:

VÌ SAO ĐẠO ĐỨC AI CẤP THIẾT HÔM NAY?

Mục tiêu học tập

- *Gồm 12 mục tiêu (Learning Objectives-LO) như sau:*

Nhận biết

- *Nhận biết vì sao đạo đức AI không còn là vấn đề học thuật, mà là yêu cầu thực tiễn trong vận hành doanh nghiệp hiện nay*
- *Hiểu được các lực đẩy chính khiến rủi ro AI gia tăng, gồm mở rộng quy mô, tự động hóa quyết định, và phụ thuộc ngày càng lớn vào dữ liệu*
- *Phân biệt được rủi ro kỹ thuật của mô hình với rủi ro xã hội, pháp lý và quản trị do AI gây ra*
- *Nhận diện các nhóm tác hại điển hình của AI trong doanh nghiệp, như thiên kiến, xâm phạm quyền riêng tư, thiếu minh bạch, mất an toàn, và đứt gãy trách nhiệm*

Năng lực ra quyết định

- *Đánh giá mức độ ảnh hưởng của một hệ thống AI đối với con người, tổ chức và xã hội, để xác định mức độ ưu tiên kiểm soát*
- *Phân tích một sự cố AI theo chuỗi nguyên nhân quản trị, thay vì quy giản về lỗi thuật toán hoặc dữ liệu*

- *Xác định được các tình huống sử dụng AI có rủi ro cao cần khung đạo đức và giám sát chặt chẽ ngay từ đầu*
- *Kết nối đạo đức AI với lợi ích kinh doanh, bao gồm giảm rủi ro pháp lý, bảo vệ uy tín, và tăng niềm tin của khách hàng và đối tác*

Năng lực truyền thông nội bộ

- *Trình bày ngắn gọn và nhất quán lý do tổ chức cần quan tâm đến đạo đức AI với lãnh đạo, đồng nghiệp và đối tác*
- *Phân biệt được các cách hiểu sai phổ biến về đạo đức AI, như "chỉ là tuân thủ pháp luật" hoặc "chỉ là vấn đề kỹ thuật"*
- *Hình thành một lập trường ban đầu về đạo đức AI cho tổ chức hoặc dự án, làm nền cho các chương tiếp theo*
- *Đặt được các câu hỏi đúng để dẫn sang các nội dung chuẩn hóa thuật ngữ, vòng đời hệ thống và khung quản trị AI*

1. CÂU CHUYỆN MỞ ĐẦU

Sáng thứ Hai, anh Tuấn, tài xế xe ôm công nghệ ở TP. Hồ Chí Minh, đứng trước quầy "một cửa" với một nỗi lo rất đời thường, giấy phép lái xe sắp hết hạn. Anh đã thử làm thủ tục cấp đổi trực tuyến cho nhanh, đăng nhập Cổng Dịch vụ công bằng tài khoản VNeID, nhưng hệ thống liên tục báo lỗi kiểu "các thông tin cập nhật thất bại do thông tin không khớp với thông tin tại cơ sở dữ liệu (CSDL) Dân cư", rồi dừng lại ở một nút "liên hệ nơi cư trú để cập nhật dữ liệu dân cư". Anh bấm đi bấm lại, cập nhật lại thông tin định danh, vẫn thế.

Chị Hạnh, cán bộ tiếp nhận, nhìn hồ sơ điện tử trên màn hình và nói nhỏ, "Không phải thiếu giấy tờ, mà là dữ liệu không đối soát được. Bên em không sửa trực tiếp trên hệ thống quốc gia. Anh phải về Công an nơi cư trú để chỉnh thông tin dân cư, rồi quay lại cập nhật từ CSDL quốc gia về dân cư." Chị chỉ anh một quy trình rất đúng, nhưng nghe vẫn như một mê cung, gọi tổng đài, gửi email, đính kèm ảnh hai mặt căn cước, chờ phản hồi.

Anh Tuấn gằn giọng, "Tôi chạy xe kiếm cơm từng ngày, thiếu giấy tờ là dừng chạy. Sai dữ liệu thì ai chịu trách nhiệm? Tôi phải chứng minh mình

là mình bằng cách nào?" Chị Hạnh không tranh cãi, chỉ thở dài, "Em hiểu, nhưng hệ thống đang 'siết' để chống gian lận, nên nó khóa theo đối soát."

Một kỹ sư công nghệ thông tin được gọi xuống hỗ trợ. Anh kỹ sư mở nhật ký (log), giải thích thẳng, "Vướng phổ biến là dữ liệu cấp bằng lái trước đây, dùng CMND 9 số hoặc bằng giấy, không trùng khớp với dữ liệu dân cư hiện tại. Khi hai kho dữ liệu không đồng bộ, đối sánh rớt, người dùng chịu trận." Chị Hạnh nhìn bạn, hỏi một câu rất khó, "Vậy trong lúc dữ liệu chưa sạch, ai là người giải thích cho dân, ai là người cho họ quyền sửa sai, quyền khiếu nại, quyền được phục vụ?"

Trưa hôm đó, phòng họp nóng lên vì một đoạn clip bắt đầu lan trên mạng. Trưởng phòng nói đúng trọng tâm, "Vấn đề không chỉ là lỗi kỹ thuật. Đây là chuyện trách nhiệm giải trình, khi một quyết định tự động gây ảnh hưởng tới sinh kế của người dân, nhưng không ai đứng ra giải thích được." Anh Tuấn không cần biết tính năng ưu việt của hệ thống sử dụng dữ liệu lớn (Big Data) của Trí tuệ nhân tạo (AI), nhưng cái anh đang gặp chính là hệ quả của hệ thống tối ưu theo rủi ro và đối soát dữ liệu, trong bối cảnh dữ liệu liên thông còn lệch, giữa cơ sở dữ liệu dân cư và dữ liệu giấy phép lái xe.

Câu hỏi mở ra chương này không phải "AI có thông minh không?", mà là "ta đã thiết kế cơ chế để con người hiểu, kiểm chứng, sửa sai và quy trách nhiệm cho quyết định tự động chưa?"

2. LUẬN ĐIỂM TRUNG TÂM

Luận điểm trung tâm của chương này là, đạo đức AI không bắt đầu từ triết lý cao siêu, nó bắt đầu từ những điểm nghẽn rất "đời", dữ liệu lệch, quyết định tự động, người dân hoặc khách hàng không biết hỏi ai, và tổ chức không biết phải chịu trách nhiệm ở lớp nào. Khi AI đi vào vận hành, đặc biệt trong dịch vụ công, "đúng mô hình" chưa đủ, hệ thống phải "đúng quy trình", có đường giải thích, có quyền phản hồi, có cơ chế sửa sai, và có người đứng tên chịu trách nhiệm giải trình (accountability) thay vì núp sau câu "lỗi hệ thống công nghệ thông tin".

Chương này hứa với bạn ba việc rất cụ thể. Một là, bạn sẽ có một "ngôn ngữ chung" để nói về AI có trách nhiệm (responsible AI) trong tổ chức,

gồm các khái niệm cốt lõi như công bằng (fairness), minh bạch, khả năng giải thích (explainability), trách nhiệm giải trình, quyền riêng tư (privacy), an toàn (safety) và an ninh (safety and security), và khả năng khiếu nại (contestability). Hai là, bạn sẽ biết cách chuyển các khái niệm đó thành cơ chế kiểm soát, ví dụ checklist yêu cầu, tiêu chí chấp nhận, ghi nhật ký (logging), chỉ số cần theo dõi, và điểm kiểm tra trong quy trình. Ba là, bạn sẽ có một công cụ dùng ngay trong 10–15 phút để rà soát một tình huống sử dụng AI, xem chỗ nào đang gặp vấn đề, chỗ nào đang "đứt trách nhiệm", và chỗ nào cần bổ sung kiểm soát trước khi đưa vào vận hành.

Trong cuốn sách, Chương 1 là lời cảnh tỉnh dựa trên dẫn chứng cụ thể, giúp người đọc nhận ra vì sao "đạo đức AI" là một yêu cầu trong vận hành, chứ không chỉ là khẩu hiệu. Chương 2 sẽ đi sâu vào hệ thuật ngữ, các khái niệm cốt lõi và khung tham chiếu để cả tổ chức nói cùng một thứ tiếng, không nói lẫn lộn giữa "độ chính xác" và "đáng tin cậy", giữa "tính năng" và "nghĩa vụ". Sang Chương 3 và Chương 4, ta sẽ nối ngôn ngữ đó với nền tảng triết học đạo đức và vòng đời hệ thống AI, để bạn biết đặt kiểm soát đúng chỗ, đúng lúc, đúng người.

3. LA BÀN NỘI DUNG

3.1. Đạo đức AI từ học thuật sang vận hành

Trong nhiều năm, đạo đức AI thường được đề cập tới như một câu chuyện "đúng, sai" trong hội thảo, hoặc một bộ nguyên tắc treo tường để làm hình ảnh. Nhưng khi AI đi vào quy trình thật, nó chạm vào quyết định thật, tiền thật, quyền lợi thật, và đôi khi là sinh kế của con người. Lúc đó, đạo đức AI không còn là lý thuyết thuần túy, nó trở thành một yêu cầu của vận hành, vì chỉ cần một quyết định tự động sai hoặc thiếu giải thích, tổ chức sẽ trả giá bằng khiếu nại, mất niềm tin, rủi ro pháp lý, và chi phí xử lý sự cố. Mục này gắn với LO1 và LO4, nhằm giúp bạn thấy vì sao đạo đức AI đã là "công việc của vận hành", không phải chuyện phụ trợ.

Điểm quan trọng là, "tác hại" của AI không chỉ là sai số dự đoán. Một hệ thống có thể đạt độ chính xác cao mà vẫn gây hại, chẳng hạn như chứa đựng các thiên kiến khiến tổ chức vô tình phân biệt đối xử dù vận hành đúng kỹ thuật (National Institute of Standards and Technology (NIST),

2023). Nó có thể xâm phạm quyền riêng tư khi thu thập, suy diễn các thông tin nhạy cảm từ dữ liệu rời rạc, hoặc mất an toàn và an ninh (security) khi bị tấn công bằng kỹ thuật tiêm lệnh nhắc (prompt injection) để sinh ra nội dung nguy hiểm (National Institute of Standards and Technology (NIST), 2024). Nó có thể thiếu minh bạch, người dùng không biết vì sao bị từ chối, cũng không biết phải làm gì để được xem xét lại. Và nguy hiểm nhất về mặt quản trị, nó có thể làm đứt gãy trách nhiệm (accountability break), mọi người đẩy nhau, rồi kết luận bằng câu "máy tính ra kết quả" (Floridi, 2023).

Bạn sẽ thấy điều này rất rõ qua các tình huống quen thuộc. Một mô hình tuyển dụng tự động loại ứng viên vì học vấn hoặc địa chỉ, nhưng thực ra nó đang dùng các "tín hiệu thay thế" (proxy signals) cho giới tính, vùng miền, hoặc tầng lớp, dẫn đến loại nhầm người phù hợp (Barocas et al., 2023). Một chatbot chăm sóc khách hàng trả lời quá nhiệt tình, vô tình lộ thông tin đơn hàng, số điện thoại, hoặc dữ liệu tài khoản khi bị hỏi vòng vèo. Một hệ thống chống gian lận chặn nhầm khách hàng tốt vì hành vi khác thường trong một ngày, và khách hàng đó không nhận được lời giải thích hay một kênh khiếu nại rõ ràng. Ba ví dụ này cho thấy, lỗi không nằm ở một chỉ số đo lường (metric), mà nằm ở cách AI được đặt vào quy trình, đặt vào quyền lực quyết định, và đặt vào mối quan hệ niềm tin với con người.

Mini story 1: Khi AI tuyển dụng "học" cách phân biệt đối xử

La bàn nội dung: Đạo đức AI vận hành

Khái niệm cốt lõi: Tác hại không đồng nghĩa với sai số, Thiên kiến đại diện (proxy bias)

Năm 2014, Amazon, một trong những tập đoàn công nghệ lớn nhất thế giới, bắt đầu xây dựng một công cụ AI để tự động hoá việc sàng lọc hồ sơ ứng viên (Dastin & J, 2018). Mục tiêu rất rõ ràng, tìm ra những nhân tài hàng đầu từ hàng núi hồ sơ một cách nhanh nhất. Hệ thống được huấn luyện dựa trên dữ liệu tuyển dụng trong khoảng 10 năm trước đó. Về mặt kỹ thuật, mô hình "làm tốt việc được giao", nó học các mẫu hình của những người từng được tuyển.

Tuy nhiên, đến khoảng năm 2015, các kỹ sư phát hiện ra một vấn đề nghiêm trọng, hệ thống chấm điểm bất lợi cho ứng viên nữ (Dastin & J, 2018). Vì dữ liệu lịch sử trong lĩnh vực này chủ yếu là nam, mô hình có xu hướng coi các tín hiệu gắn với "nam giới" là tín hiệu của "thành công". Nó học các mẫu hình của những người từng được tuyển, từ đó hạ điểm hồ sơ có chứa từ "women's", ví dụ "women's chess club captain", đội trưởng đội cờ vua nữ, và hạ thấp xếp hạng của ứng viên từ các trường đại học nữ .

Đội dự án đã thử nhiều cách sửa, ví dụ yêu cầu mô hình "trung tính" với các từ khoá giới tính cụ thể. Nhưng bài toán khó nằm ở chỗ, khi bạn bịt một cửa, mô hình vẫn có thể tìm biến thay thế, tức "tín hiệu đại diện" (proxy signals), từ cách dùng ngôn ngữ đến các mẫu từ vựng thường xuất hiện trong hồ sơ nam giới. Nói gọn, nó không cần nhìn "giới tính" trực tiếp, nó chỉ cần nhìn một bóng dáng của giới tính trong dữ liệu.

Cuối cùng, dự án bị dừng, nhóm bị giải tán vào khoảng đầu năm 2017 vì không thể bảo đảm công bằng (Dastin & J, 2018). (Reuters cũng ghi nhận Amazon cho biết công cụ này không được dùng như một hệ thống ra quyết định tuyển dụng trực tiếp.) Bài học ở đây là, một hệ thống có thể tối ưu rất tốt về mặt toán học, tìm người giống "nhân viên cũ" nhất, nhưng lại thất bại về mặt đạo đức vận hành, vì nó tái tạo và khuếch đại định kiến xã hội. Đây là ví dụ điển hình cho thấy "tác hại" (harms) trong vận hành thực tế khác xa với "sai số" (errors) trong phòng thí nghiệm. Điểm cốt lõi ở đây là: lỗi không nằm ở phép tính, lỗi nằm ở thước đo.

Tài liệu tham khảo

1. Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/
2. Reuters. (2018, October 10). Amazon ditched AI recruiting tool that favored men for technical jobs. The Guardian. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Lỗi hiểu sai phổ biến nhất là coi đạo đức AI như một lớp "trang điểm" cho thương hiệu, hoặc nghĩ "cứ đúng luật là xong". Đúng luật là cần, nhưng chưa đủ, vì nhiều rủi ro vận hành xảy ra ở vùng xám trước khi luật kịp gọi tên, hoặc ở chỗ luật không thể mô tả hết, như khả năng giải thích, khả năng khiếu nại, hay cơ chế phân quyền trách nhiệm khi quyết định là bán tự động. Nếu bạn coi đạo đức AI là chuyện hình ảnh, bạn sẽ thiếu nhật ký và ghi vết, thiếu kiểm thử, thiếu quy trình phản hồi, và thiếu người chịu trách nhiệm, rồi đến lúc xảy ra sự cố thì không có gì để chứng minh mình đã làm đúng.

Khi AI đã là vận hành, ta không thể chỉ hỏi "AI có tốt không?". Ta phải hỏi vì sao rủi ro đang tăng nhanh? những lực đẩy nào làm rủi ro phình ra theo cấp số nhân, và tổ chức đang bị kéo vào thế bị động ra sao nếu không có ngôn ngữ và cơ chế kiểm soát ngay từ đầu?

3.2. Ba lực đẩy làm rủi ro AI tăng nhanh

Có những lực đẩy mang tính hệ thống, khiến rủi ro AI bùng lên trong tổ chức ngay cả khi "ý định ban đầu" rất tốt. Nếu chỉ nhìn rủi ro như lỗi

kỹ thuật lẻ tẻ, các biện pháp xử lý luôn mang tính tình thế, vì nguyên nhân nằm ở cách AI được đưa vào vận hành ở quy mô lớn, với mức tự động hóa cao, và phụ thuộc dữ liệu ngày càng sâu (National Institute of Standards and Technology (NIST), 2023).

Lực đẩy thứ nhất là mở rộng quy mô (scale). Một sai lệch nhỏ ở mức cá nhân, khi nhân lên hàng trăm nghìn hoặc hàng triệu quyết định, sẽ thành tác hại lớn về kinh tế và niềm tin. Ở quy mô này, "lỗi nhỏ" không còn là lỗi nhỏ, nó trở thành sai số có hệ thống. Một rule gợi ý sản phẩm chỉ cần sai một tham số, có thể làm lệch hành vi mua sắm của cả một nhóm khách hàng trong nhiều tuần, kéo theo tồn kho, chi phí marketing, và cảm giác bị thao túng. Khi tổ chức chạy trên bảng điều hành (dashboard) và KPI, hệ thống sẽ tiếp tục "đẩy mạnh cái sai" nếu cái sai đó vô tình làm đẹp chỉ số ngắn hạn.

Lực đẩy thứ hai là tự động hóa (automation) và tốc độ. AI loại bỏ độ trễ của con người, nhưng cũng loại bỏ luôn các "khoảng dừng" để suy xét. Rủi ro lớn nhất là tâm lý "cài đặt và quên" (set and forget), khi người vận hành tin tưởng mù quáng vào hệ thống tự động và mất khả năng kiểm soát khi sự cố xảy ra. Các quy định pháp lý hiện đại bắt buộc phải duy trì cơ chế giám sát của con người để ngăn chặn rủi ro tự động hóa quá mức này (Regulation (EU) 2024/1689 on Artificial Intelligence, 2024).

Lực đẩy thứ ba là phụ thuộc dữ liệu (data dependence). AI học từ dữ liệu, vận hành bằng dữ liệu, và bị trói bởi dữ liệu. Khi dữ liệu sai, lệch, thiếu đại diện, hoặc không đồng bộ giữa các hệ thống, nó trở thành "luật ngầm" mà mô hình tuân theo, rất khó phát hiện nếu bạn không có cơ chế kiểm chứng. Tệ hơn, nhiều tổ chức coi dữ liệu như sự thật khách quan, trong khi dữ liệu thường là dấu vết của quá khứ, gồm cả thiên kiến, chính sách cũ, và sai sót nhập liệu; khi AI dựa vào dữ liệu để tối ưu, nó có thể vô tình hợp thức hóa các lệch lạc đó, biến một lỗi dữ liệu thành một quyết định có vẻ "khoa học" (Barocas et al., 2023).

Mini story 2: Robodebt - Khi tự động hóa tước đoạt quyền con người

La bàn nội dung: Ba lực đẩy làm rủi ro AI tăng nhanh

Khái niệm cốt lõi: Tự động hóa (Automation), Đứt gãy trách nhiệm (Accountability break)

Từ năm 2015 đến khoảng 11/2019, chính phủ Úc triển khai hệ thống "Robodebt" nhằm tự động phát hiện và thu hồi các khoản trợ cấp xã hội bị trả thừa. Thay vì quy trình cũ có con người kiểm tra, hệ thống mới dùng thuật toán đối soát dữ liệu thu nhập từ cơ quan thuế với dữ liệu báo cáo của người nhận trợ cấp. Một cơ chế gây tranh cãi là "bình quân thu nhập" (income averaging), suy diễn rằng nếu tổng thu nhập năm khác với báo cáo theo kỳ, hệ thống tự coi đó là nợ và gửi thông báo đòi tiền.

Hậu quả là nặng nề. Hàng trăm nghìn người dân, ở quy mô được báo cáo là trên 400.000 trường hợp, nhận được các thông báo nợ sai. Hệ thống đã làm lệch một nguyên tắc pháp lý quan trọng, thay vì cơ quan nhà nước phải chứng minh người dân sai, người dân bị đẩy vào thế phải tự chứng minh mình không nợ, tức "đảo ngược gánh nặng chứng minh" (reversed onus of proof), trước một quy trình máy móc mà họ không hiểu và rất khó đối thoại. Nhiều người rơi vào khủng hoảng tài chính, suy sụp tinh thần, và các nguồn báo cáo có nêu những trường hợp tự tử được cho là có liên quan đến áp lực truy thu kéo dài (Karp & P, 2023).

Sau đó, chính phủ Úc thừa nhận sai sót ở quy mô lớn, và chương trình trở thành một vụ việc quốc gia với các dàn xếp, hoàn trả ở cỡ hàng tỉ đô la Úc (AUD), cùng điều tra cấp Ủy ban Hoàng gia (Royal Commission). Robodebt là minh chứng đau đớn cho "lực đẩy tự động hoá" (automation) khi thiếu giám sát của con người. Nó tạo ra một "đứt gãy trách nhiệm" (accountability break), nơi hệ thống tự phát sinh quyết định ảnh hưởng trực tiếp đến sinh kế, nhưng trách nhiệm giải trình bị kéo giãn đến mức "không ai là người ra quyết định". Và khi quyền phản đối (right to contest), quyền khiếu nại không được thiết kế như một phần mặc định của hệ thống, người yếu thế sẽ là người trả giá đầu tiên (Royal Commission into the Robodebt Scheme, 2023).

Đây không phải lỗi của một thuật toán đơn lẻ, mà là hệ quả của cả một chuỗi lựa chọn quản trị. Bối cảnh và mức ảnh hưởng quyết định bạn phải nghiêm đến đâu, phúc lợi xã hội không phải chỗ để "tự động hoá cho nhanh".

Tài liệu tham khảo

1. Karp, P. (2023, July 7). Robodebt royal commission final report: what did it find and what will happen next?. The Guardian. https://www.theguardian.com/australia-news/2023/jul/07/robodebt-royal-commission-final-report-what-did-it-find-and-what-will-happen-next
2. Royal Commission into the Robodebt Scheme. (2023, July 7). Report of the Royal Commission into the Robodebt Scheme. Commonwealth of Australia. https://robodebt.royalcommission.gov.au/

Lỗi hiểu sai thường gặp có hai dạng. Một là nghĩ "đã có AI thì mọi thứ khách quan hơn", trong khi AI chỉ khách quan nếu dữ liệu và mục tiêu tối ưu được thiết kế để giảm thiên lệch, và có kiểm soát liên tục. Hai là tin rằng "càng nhiều dữ liệu càng đúng", trong khi dữ liệu nhiều nhưng bẩn hoặc lệch có thể làm mô hình tự tin hơn vào cái sai, và gây hại rộng hơn. Nhiều dữ liệu không thay thế được dữ liệu phù hợp, có nguồn gốc rõ ràng, và được kiểm định.

Nếu rủi ro tăng vì cơ chế vận hành, ta phải biết phân biệt rủi ro kỹ thuật của mô hình với rủi ro xã hội, pháp lý, và quản trị của hệ thống. Vì chỉ khi phân biệt đúng, bạn mới đặt đúng kiểm soát, đúng người chịu trách nhiệm, và đúng chỉ số cần theo dõi.

3.3. Hai lớp rủi ro, kỹ thuật và quản trị xã hội

Cần tách bạch lỗi mô hình với lỗi hệ thống xã hội, pháp lý, và quản trị, để tránh xử lý sai nguyên nhân. Nhiều tổ chức khi gặp sự cố AI phản xạ theo bản năng kỹ thuật, gọi nhóm làm dữ liệu (data team) vào "tối ưu thêm", tăng dữ liệu, chỉnh tham số, đổi mô hình. Làm vậy đôi khi cần, nhưng nếu bạn không phân biệt hai lớp rủi ro, bạn sẽ sửa đúng phần dễ nhìn mà bỏ qua phần gây hại thật.

Lớp thứ nhất là rủi ro kỹ thuật (technical risks). Đây là các vấn đề thuộc bản thân mô hình và pipeline, ví dụ quá khớp (overfitting) khiến mô hình đẹp trên tập dữ liệu kiểm tra (test set) nhưng rớt ngoài đời, trôi dạt dữ liệu (data drift), và trôi dạt khái niệm (concept drift) khiến hiệu năng giảm theo thời gian, ảo giác (hallucination) của AI tạo sinh (Generative AI, GenAI) tạo ra thông tin bịa mà nghe rất thuyết phục, độ mạnh và ổn định (robustness) kém khiến chỉ cần thay đổi nhỏ ở đầu vào là kết quả đổi lớn, hoặc mô hình dễ bị tấn công theo kiểu tiêm lệnh nhắc, ngộ độc dữ liệu (data poisoning). Đây là lớp rủi ro mà kỹ sư thường quen tay xử lý, bằng kiểm thử, giám sát, tái huấn luyện, kiểm soát dữ liệu, và gia cố hệ thống (National Institute of Standards and Technology (NIST), 2024).

Lớp thứ hai là rủi ro xã hội và quản trị (socio technical and governance risks). Loại rủi ro này xuất hiện khi AI bước vào quan hệ quyền lực và lợi ích giữa con người với con người, giữa tổ chức với người dùng, và giữa tổ

chức với pháp luật. Ở lớp này, có thể xuất hiện hiện tượng phân biệt đối xử gián tiếp (proxy bias), mô hình không hề dùng biến "giới tính" hay "vùng miền" nhưng lại dùng tín hiệu thay thế như trường học, địa chỉ, lịch sử giao dịch, khiến một nhóm bị loại trừ có hệ thống. Bên cạnh đó là tình trạng thiếu giải thích (lack of explainability), người bị ảnh hưởng không biết vì sao bị từ chối, không biết phải làm gì để được xem xét lại. Một vấn đề khác "không có cơ chế khiếu nại" (no contestability), khiến các quyết định tự động trở thành cánh cửa đóng kín. Và cuối cùng là câu hỏi trách nhiệm giải trình, ai chịu trách nhiệm khi hệ thống gây hại? người ký phê duyệt, người vận hành, hay nhà thầu, hay "không ai cả"? (Regulation (EU) 2024/1689 on Artificial Intelligence, 2024).

Điểm khó là một mô hình có thể "đúng kỹ thuật" mà vẫn gây loại trừ hoặc bất công. Ví dụ một hệ thống chấm điểm rủi ro để ưu tiên xử lý hồ sơ có thể đạt AUC tốt, ổn định, không drift, không overfit, nhưng mục tiêu tối ưu lại sai, nó tối ưu giảm gian lận bằng mọi giá, nên sẵn sàng hy sinh tỷ lệ từ chối nhầm với người dùng hợp lệ. Nếu không có cơ chế khiếu nại và sửa sai, người bị từ chối nhầm sẽ chịu thiệt, còn tổ chức thì tạo ra một kiểu "bất công có hệ thống" mà báo cáo kỹ thuật không bao giờ gọi tên. Hoặc một chatbot nội bộ có hệ thống bảo vệ (guardrail) kỹ, nhưng quy trình cấp quyền truy cập dữ liệu lỏng, ai cũng hỏi được dữ liệu nhạy cảm, vậy rủi ro không nằm ở mô hình, mà nằm ở quản trị truy cập và phân quyền.

Lỗi hiểu sai thường gặp là cho rằng "cải thiện thuật toán là đủ", trong khi nhiều sự cố là sự cố quy trình, sự cố quyền lực, và sự cố trách nhiệm. Một lỗi hiểu sai khác là tin rằng "đưa thêm vài biến là hết thiên kiến". Thực tế, độ tin cậy kỹ thuật cao không đồng nghĩa với sự an toàn của hệ thống, vì rủi ro thường nảy sinh từ sự tương tác phức tạp giữa công cụ AI và bối cảnh sử dụng (Dobbe, 2022). Thiên kiến không chỉ là bài toán thống kê, nó là bài toán mục tiêu, dữ liệu, và cơ chế bảo vệ quyền lợi con người.

Khi đã xác định được hai lớp rủi ro này, cần có một cách đánh giá mức độ ảnh hưởng, để xác định hệ thống nào cần được ưu tiên kiểm soát, kiểm soát tới mức nào, và ai phải chịu trách nhiệm ở từng lớp. Vì nguồn lực luôn có hạn, và quản trị tốt là cần phân bổ nguồn lực đúng chỗ.

3.4. Đánh giá mức độ ảnh hưởng để ưu tiên kiểm soát

Việc đánh giá mức độ ảnh hưởng của một hệ thống AI tới con người, tổ chức và xã hội là cần thiết để từ đó nhận diện tình huống sử dụng rủi ro cao và từ đó ưu tiên kiểm soát. Trong thực tế, không thể kiểm soát mọi thứ với cùng một mức chặt chẽ, vì nguồn lực luôn hữu hạn, tiến độ triển khai luôn bị thúc ép, và áp lực "phải có AI" luôn lớn. Vấn đề không phải làm nhiều hay ít, mà là xác định đúng những điểm cần kiểm soát chặt, kiểm soát sâu và có bằng chứng.

Nguyên tắc đầu tiên là, ảnh hưởng càng lớn thì yêu cầu về bằng chứng và giám sát càng chặt. "Bằng chứng" ở đây không chỉ là một báo cáo độ chính xác (accuracy), mà là bằng chứng hệ thống, mô hình có được kiểm thử trong điều kiện gần thực tế không, có giám sát trôi dạt (drift) không? có cơ chế phát hiện bất thường không, có nhật ký và ghi vết đủ để truy vết không, và có quy trình can thiệp của con người không (Prem, 2025). "Giám sát" không chỉ là nhìn bảng điều hành, mà là thiết kế vòng phản hồi vận hành, phát hiện sớm, phản ứng nhanh, sửa sai có kiểm soát, và ghi nhận đầy đủ để kiểm toán sau này. Nếu một quyết định có thể làm một cá nhân mất quyền lợi hoặc bị loại khỏi một cơ hội, thì tiêu chuẩn chứng minh phải khác hẳn so với một hệ thống chỉ gợi ý nội dung giải trí.

Nguyên tắc thứ hai là, dấu hiệu "rủi ro cao" thường nằm ở bản chất của quyết định, chứ không nằm ở công nghệ. Hãy xem xét ba câu hỏi rất thực tiễn sau: Quyết định này có liên quan trực tiếp đến quyền lợi, tiền bạc, hoặc điều kiện sống không? Quyết định này có liên quan đến an toàn, sức khỏe, hoặc tính mạng không? Quyết định này có phân bổ cơ hội, quyền tiếp cận, hoặc tạo ra loại trừ không? Chỉ cần bạn trả lời "có" cho một trong ba câu hỏi, bạn đang đứng trước một tình huống sử dụng cần kiểm soát mạnh hơn, có thể cần thẩm định, cần giải thích, cần kênh khiếu nại, và cần người chịu trách nhiệm rõ ràng. Đây là cách nhìn rất gần với logic quản trị dựa trên rủi ro (risk based approach), thay vì nhìn theo mặt lợi ích về công nghệ.

Nguyên tắc thứ ba là xếp ưu tiên theo tư duy quản trị dựa trên rủi ro, nghĩa là phân loại tình huống sử dụng theo mức ảnh hưởng và khả năng gây hại, rồi xác định mức kiểm soát tương ứng. Nếu bạn đã từng đọc cách

tiếp cận phân tầng rủi ro trong các khung quản trị và pháp lý, bạn sẽ thấy một đường dây chung, các hệ thống ảnh hưởng lớn tới quyền cơ bản và cơ hội của con người sẽ bị yêu cầu nhiều nghĩa vụ hơn, từ quản trị dữ liệu, tài liệu hóa, đến giám sát sau triển khai và cơ chế phản hồi (Smuha & Yeung, 2025). Dù bạn chưa dùng một khung nào, tư duy này vẫn là công cụ lãnh đạo rất mạnh, nó giúp bạn nói "không" với những triển khai hấp tấp ở vùng rủi ro cao, và nói "được" với những thử nghiệm an toàn ở vùng rủi ro thấp.

Các ví dụ kinh điển có thể thấy rõ trong thực tiễn. Chấm điểm tín dụng quyết định ai được vay, ai bị từ chối, qua đó trực tiếp phân bổ cơ hội tài chính. Tuyển dụng quyết định ai được vào, ai bị loại, tác động đến sinh kế và công bằng. Phân loại ưu tiên cấp cứu (triage) y tế, phân luồng ưu tiên điều trị, động tới an toàn và tính mạng. Phân loại hồ sơ hành chính, nếu dùng để tự động từ chối hoặc trì hoãn, có thể tạo ra loại trừ dịch vụ công và bất bình đẳng tiếp cận (Smuha & Yeung, 2025). Điểm chung của các tình huống sử dụng này là, chúng không chỉ tối ưu một KPI nội bộ, chúng chạm vào quyền lợi thật của con người, vì vậy yêu cầu kiểm soát phải đi trước, không phải chỉ được đặt ra sau khi xảy ra sự cố.

Lỗi hiểu sai thường gặp là coi mọi tình huống sử dụng như nhau, "tình huống sử dụng nào cũng làm nhanh" để kịp lộ trình (roadmap), hoặc "thử trước rồi tính" với cả những quyết định có ảnh hưởng lớn. Thử nghiệm giới hạn là tốt, nhưng thử nghiệm phải có biên độ an toàn. Với vùng rủi ro cao, thử nghiệm mà không có kịch bản sai, không có cơ chế can thiệp của con người (human override), không có kênh khiếu nại, và không có nhật ký và ghi vết chuẩn, thì đó không phải thử nghiệm, đó là đem người dùng thật ra làm chuột bạch (Smuha & Yeung, 2025). Một tổ chức có trách nhiệm và tầm nhìn sẽ không lựa chọn cách làm như vậy vì hậu quả có thể rất lớn.

Khi đã xác định được thứ tự ưu tiên, cần nhìn một sự cố AI như một chuỗi nguyên nhân quản trị, để không dừng lại ở việc đổ lỗi cho mô hình. Chỉ khi truy được chuỗi nguyên nhân, bạn mới thiết kế được kiểm soát đúng chỗ, và tránh lặp lại cùng một sự cố với một cái tên khác.

3.5. Mổ xẻ sự cố AI theo chuỗi nguyên nhân quản trị

Một đột phá tư duy trong quan điểm về AI là chuyển từ tư duy "lỗi kỹ

thuật" sang tư duy "đứt gãy quản trị". Khi một sự cố AI xảy ra, phản ứng bản năng của nhiều đội là tìm "lỗi" ("bug") trong mô hình, hoặc tìm một tham số để chỉnh sửa. Nhưng thực tế, đa số sự cố nghiêm trọng không bắt đầu từ mô hình mà bắt đầu từ mục tiêu sai, dữ liệu không kiểm chứng, quy trình phê duyệt lỏng, vận hành không giám sát, phản hồi không có cửa vào, và truyền thông không nhận được câu trả lời thỏa đáng. Tư duy an toàn hệ thống hiện đại chỉ ra rằng độ tin cậy của phần mềm hay mô hình không đảm bảo sự an toàn của cả hệ thống, mà tai nạn thường là kết quả của các quy trình phức tạp trong cả hệ thống kỹ thuật - xã hội (Dobbe, 2022).

Một cách nhìn thực dụng là xem sự cố như một chuỗi, thay vì một điểm. Chuỗi đó thường đi qua các mắt xích: mục tiêu tối ưu được đặt ra như thế nào, KPI nào đang dẫn dắt hành vi của nhóm triển khai, dữ liệu đến từ đâu và đã được kiểm định chưa, ai phê duyệt cho hệ thống chạy ở mức tự động hóa nào, ngưỡng chấp nhận sai số là gì, ai giám sát trôi dạt mô hình (model drift) và chất lượng đầu ra. Để quản trị điều này, hoạt động kiểm toán không chỉ là kiểm tra kỹ thuật tại một thời điểm, mà phải là một quy trình đánh giá xuyên suốt vòng đời từ phát triển đến sau khi triển khai để xác định các rủi ro thực tế (Raji, 2023). Nếu bạn chỉ nhìn vào mô hình, bạn sẽ bỏ qua các mắt xích này, trong khi chính chúng quyết định sự cố có trở thành khủng hoảng hay không.

Vì vậy, chúng ta tập luyện một thói quen đặt câu hỏi theo trục trách nhiệm. Bốn câu hỏi cốt lõi, "ai quyết?", "ai kiểm?", "ai chịu trách nhiệm?", "bằng chứng ở đâu?", là khung xương giúp bạn không bị cuốn vào tranh cãi kỹ thuật mà quên mất bài toán quản trị. Hãy lấy một tình huống rất phổ biến, chatbot tư vấn nội bộ hoặc tư vấn khách hàng trả lời sai. Nếu chỉ sửa mô hình, bạn sẽ thay lời nhắc (prompt) hoặc tinh chỉnh lại. Nhưng nhìn theo chuỗi nguyên nhân quản trị, câu hỏi sẽ khác: Có cơ chế đưa sự việc lên cấp cao hơn (escalation) không? Khi chatbot không chắc thì nó có chuyển cho người thật không, có ghi vết quyết định và ngữ cảnh hội thoại để truy vết không? Các khung quản trị quy định (regulatory ecosystem) hiệu quả đều yêu cầu phải thiết lập khả năng khiếu nại lên con người (human appeal) và hệ thống lưu trữ (archiving) để đánh giá tác động khi có thay đổi (Hudson, 2022).

Một cách hiểu sai thường gặp là "đổi model khác là xong" hoặc "đổ cho nhà cung cấp". Nhà cung cấp có thể có phần trách nhiệm, nhưng tổ chức vận hành vẫn phải chịu trách nhiệm về quyết định dùng AI ở đâu, ở mức tự động hóa nào. Bạn không thể thuê ngoài trách nhiệm giải trình. Việc đổ lỗi cho sự phức tạp của thuật toán hay nhà cung cấp thực chất là một hình thức "rửa tay trách nhiệm" (agency laundering) hoặc "né tránh đạo đức" (ethics shirking), nơi tổ chức cố tình tạo ra sự mơ hồ để chối bỏ nghĩa vụ đạo đức của mình (Floridi, 2023).

Chúng ta đã mổ xẻ sự cố theo chuỗi nguyên nhân quản trị, cần quay về câu hỏi thực dụng nhất mà lãnh đạo và doanh nghiệp luôn hỏi, đạo đức AI mang lại lợi ích kinh doanh gì? và vì sao đầu tư vào kiểm soát lại là một quyết định khôn ngoan chứ không phải chi phí "làm cho đẹp"?

3.6. Đạo đức AI như một bài toán niềm tin và lợi ích kinh doanh

Ba thứ mà lãnh đạo nào cũng quan tâm: pháp lý, uy tín, và chi phí cơ hội. Nếu bạn trình bày đạo đức AI như một bài giảng về "đúng sai", tổ chức sẽ nghe rồi để đó. Nhưng nếu bạn trình bày nó như một bài toán niềm tin, một loại tài sản vô hình nhưng quyết định tốc độ mở rộng thị trường và độ bền của thương hiệu, thì thái độ của tổ chức sẽ khác hẳn.

Lợi ích thứ nhất là giảm rủi ro pháp lý và chi phí khắc phục. Khi AI gây hại, chi phí không chỉ là tiền phạt, mà là chi phí điều tra, chi phí xử lý khiếu nại, chi phí dừng hệ thống, chi phí sửa dữ liệu, và chi phí cơ hội vì sản phẩm bị chậm lại. Tệ hơn, nếu không có nhật ký và ghi vết, không có tài liệu hóa, và không có quy trình phản hồi, tổ chức không chứng minh được mình đã làm đúng, kể cả khi "ý định" là tốt. Đạo đức AI ở đây là thiết kế cơ chế kiểm soát từ đầu, như các công cụ giám sát và báo cáo, để khi có sự cố bạn có thể truy vết, khoanh vùng, sửa sai có bằng chứng, và giảm thời gian gián đoạn của dự án (Prem, 2025).

Lợi ích thứ hai là giảm rủi ro truyền thông và mất uy tín. Trong thời đại mạng xã hội phát triển mạnh, sự cố AI rất dễ biến thành câu chuyện đạo đức, "máy móc vô cảm", "công ty coi thường khách hàng", "nhà nước hành dân", dù gốc rễ ban đầu có thể chỉ là một lỗi dữ liệu hoặc một quyết định tối ưu sai. Nếu tổ chức không có cách giải thích đơn giản, không có cơ chế

khiếu nại, và không có thái độ nhận trách nhiệm, khủng hoảng sẽ bùng lên nhanh hơn năng lực xử lý. Nhiều đội kỹ thuật nghĩ rằng "chúng tôi sửa rồi" là xong, nhưng truyền thông không quan tâm bạn sửa thế nào, họ quan tâm bạn đối xử với người bị ảnh hưởng ra sao. Đạo đức AI giúp bạn chuẩn bị trước các kịch bản rủi ro quan hệ công chúng (public relations risks) để bảo vệ uy tín thương hiệu.

Lợi ích thứ ba là tăng niềm tin của khách hàng, đối tác, và cơ quan quản lý. Niềm tin không đến từ demo, nó đến từ khả năng chứng minh. Khi bạn muốn mở rộng thị trường, đặc biệt là các thị trường có yêu cầu tuân thủ cao, bạn sẽ bị hỏi những câu rất cụ thể, dữ liệu đến từ đâu, có nhất quán không, có đánh giá rủi ro không, có kiểm thử công bằng không, có giám sát sau triển khai không, có cơ chế khiếu nại không, ai chịu trách nhiệm? Thực tế cho thấy sự thâm hụt niềm tin (trust deficit) đang là rào cản lớn, và việc tích hợp đạo đức không chỉ là mong muốn mà là yếu tố sống còn để mở khóa tiềm năng kinh tế số (Bansal, 2025).

Ví dụ thực dụng nhất là một sản phẩm AI trong doanh nghiệp muốn mở rộng từ một thử nghiệm (pilot) nội bộ sang dùng cho khách hàng hoặc mở sang tỉnh thành khác, quốc gia khác. Ở giai đoạn thử nghiệm, bạn có thể "đỡ" bằng con người, có thể giải thích bằng miệng, có thể sửa bằng tay. Nhưng khi mở rộng, bạn phải chứng minh hệ thống kiểm soát tồn tại và vận hành được ở quy mô, có tiêu chí chấp nhận rõ, có dấu vết kiểm toán, có quy trình xử lý sai, và có người chịu trách nhiệm. Việc thiết lập khả năng truy vết (traceability) để quy trách nhiệm đạo đức là điều kiện tiên quyết để hệ thống vận hành minh bạch trong các cấu trúc kỹ thuật xã hội phức tạp.

Lỗi hiểu sai phổ biến là nói "đạo đức làm chậm đổi mới". Thực tế, đạo đức AI làm chậm một số quyết định vội vàng, nhưng đổi lại nó giảm chi phí trả giá về sau, giảm thời gian chữa cháy, và tăng tốc mở rộng bền vững. Đạo đức AI giống như phanh của một chiếc xe mạnh, phanh tốt không làm xe chậm mà làm bạn dám chạy nhanh hơn vì bạn biết mình dừng được. Đạo đức mềm (soft ethics) mang lại lợi thế kép (dual advantage): vừa là chiến lược quản trị rủi ro để tránh những sai lầm tốn kém, vừa là chiến lược cơ hội để nắm bắt các giá trị xã hội bền vững (Floridi, 2023).

Bạn đã thấy lợi ích, bước khó tiếp theo là thuyết phục nội bộ, và tránh các hiểu nhầm phổ biến về đạo đức AI, để nó không biến thành một phong trào hình thức hoặc một "phòng ban kiểm duyệt" làm tắc mọi thứ.

3.7. Đồng thuận nội bộ, lập trường ban đầu, và câu hỏi đúng để đi tiếp

Trong tổ chức, đạo đức AI thường thất bại vì hai lý do, một là thảo luận kéo dài mà không gắn với công việc cụ thể, hai là mỗi phòng ban hiểu vấn đề theo một cách khác nhau, khiến dự án AI biến thành một cuộc "dịch thuật" liên miên giữa kỹ thuật, pháp chế, vận hành, và truyền thông. Điểm nghẽn này không nằm ở việc thiếu công nghệ, mà nằm ở việc thiếu một "nền tảng kiến thức" (knowledge foundation) để các nhóm chuyên môn thật sự trao đổi được với nhau, và nhận diện được điểm can thiệp trong vòng đời hệ thống. (Greene, 2022)

Vì vậy, bạn cần một lập trường ban đầu đủ rõ ràng để tạo đồng thuận, và một bộ câu hỏi đúng để biến đồng thuận thành hành động. Trước hết là "thông điệp 30 giây", nói cho lãnh đạo, cho đồng nghiệp, và cho đối tác. Bạn có thể dùng một câu như sau: "Chúng ta dùng AI để tăng hiệu quả và chất lượng, nhưng chỉ triển khai khi có bằng chứng kiểm soát rủi ro, có cơ chế giải thích, có kênh phản hồi và sửa sai, và có người chịu trách nhiệm rõ ràng." Tiếp theo là làm rõ ba hiểu nhầm phổ biến, vì nếu không bóc tách, bạn sẽ bị kéo lệch ngay từ buổi khởi động (kick off).

Từ đó, bạn hình thành "lập trường ban đầu" của tổ chức, thường là một trang ngắn kiểu "quan điểm đạo đức" (AI ethics stance). Cốt lõi của trang này là "định vị quy phạm" (normative positioning), tức là xác định tổ chức phải nói rõ giới hạn không được vượt qua và mức đánh đổi chấp nhận được, và những điều gì không làm, trước khi triển khai dữ liệu và tự động hóa. Nếu lập trường này không được chia sẻ đủ rộng, mọi quyết định sau đó sẽ bị giằng xé giữa lợi ích ngắn hạn và rủi ro dài hạn. (Erk, 2025) Nhưng lập trường tốt cũng không cản trở đổi mới, nó đặt "đường ray" để đổi mới không trật bánh, cho phép thử nghiệm trong vùng rủi ro thấp, có sandbox, có giới hạn tác động, có con người trong vòng lặp (human in the loop) và có tiêu chí dừng rõ ràng.

Sau cùng là bộ câu hỏi để tiếp tục triển khai, vì câu hỏi đúng sẽ dẫn bạn sang chuẩn hóa thuật ngữ, đặt đạo đức vào vòng đời hệ thống, và chọn khung quản trị. Câu hỏi đặt ra là: hệ thống này tác động ai, quyền lợi nào, mức ảnh hưởng đến đâu, dữ liệu đến từ đâu và có hợp pháp, hợp lệ không, mô hình có thể giải thích ở mức nào cho từng nhóm đối tượng, có cơ chế khiếu nại và sửa sai không, ai phê duyệt mức tự động hóa và ngưỡng sai chấp nhận được, nhật ký và ghi vết, dấu vết kiểm toán cần gì, chỉ số nào cần theo dõi sau triển khai, và ai chịu trách nhiệm cuối cùng khi có thiệt hại. Lỗi hiểu sai thường gặp là biến đạo đức AI thành khẩu hiệu, nói rất hay nhưng không có câu hỏi vận hành cụ thể, không có tiêu chí chấp nhận, không có quy trình giám sát, và không có người chịu trách nhiệm. Khẩu hiệu chỉ làm mọi người yên tâm giả, đến lúc sự cố xảy ra thì không có gì để bám, vừa không bảo vệ được người dùng, vừa không bảo vệ được tổ chức. Nhìn đúng, đạo đức không chỉ là rào cản, nó vừa là chiến lược cơ hội, vừa là giải pháp quản trị rủi ro để tránh những sai lầm tốn kém (Floridi, 2023).

3.8. Tiếu kết

Chương 1 khẳng định một chuyển dịch căn bản: đạo đức AI không còn là phần "đúng sai" trên sân khấu học thuật, mà là kỷ luật vận hành trong hệ thống thật, nơi AI chạm vào quyền lợi, niềm tin, và rủi ro pháp lý. Khi AI mở rộng quy mô, tăng tự động hóa, và phụ thuộc dữ liệungày càng sâu, tác hại không còn nằm ở một chỉ số sai số, mà nằm ở cách mục tiêu tối ưu, dữ liệu, quy trình, và quyền lực quyết định gắn vào nhau. Vì thế, cần phân biệt rõ rủi ro kỹ thuật với rủi ro xã hội và quản trị, rồi dùng đánh giá mức độ ảnh hưởng để ưu tiên kiểm soát, thay vì chữa cháy theo phản xạ "tối ưu mô hình". Cách phân tích sự cố theo chuỗi nguyên nhân quản trị cũng nhắc lại một thực tế quan trọng: tổ chức không thể thuê ngoài trách nhiệm giải trình, muốn quản trị hiệu quả thì phải có cơ chế bằng chứng, từ nhật ký và ghi vết đến quy trình phản hồi và người chịu trách nhiệm.

Từ đây một câu hỏi xuất hiện: nếu đạo đức AI đã là vận hành, thì "ngôn ngữ" nào giúp các nhóm kỹ thuật, pháp chế, kinh doanh, vận hành nói cùng một thứ tiếng để ra quyết định nhất quán? Đó là lý do chương 2 đi vào ngôn ngữ chung, tách ba lớp giá trị (values), nguyên tắc (principles), yêu cầu

triển khai (requirements), và chốt thuật ngữ để tranh luận không lệch pha. Và khi đã có ngôn ngữ chung, chương 3 sẽ cung cấp "cú pháp lập luận" của triết học đạo đức (moral philosophy), ba lăng kính đức hạnh luận (virtue ethics), bổn phận luận (deontology), hệ quả luận (consequentialism), để biến tranh luận thành quyết định có thể biện minh, có điều kiện chấp nhận, và có nền cho chính sách.

4. CÁC KHÁI NIỆM CỐT LÕI

Đây là 5 khái niệm dùng để ra quyết định, không phải lý thuyết suông:

4.1. Đạo đức AI vận hành

Định nghĩa: Đạo đức AI vận hành là tập nguyên tắc, vai trò, quy trình, và bằng chứng vận hành, giúp giảm tác hại và tăng khả năng chịu trách nhiệm khi AI tham gia hoặc ảnh hưởng đến quyết định. "Bằng chứng" ở đây là những thứ kiểm được, như nhật ký, biên bản phê duyệt, kết quả kiểm thử, tài liệu dữ liệu, và quy trình xử lý khiếu nại.

Vì sao quan trọng: Nếu coi đạo đức là "phụ kiện", bạn sẽ chỉ thấy rủi ro khi đã bị khách hàng phản ứng, bị cơ quan quản lý hỏi, hoặc phải dừng hệ thống để điều tra. Lúc đó bạn không chỉ thiếu câu trả lời, bạn thiếu cả chứng cứ để tự bảo vệ.

Ví dụ: Dự án AI tạo sinh chăm sóc khách hàng có quy định dữ liệu nào được dùng, ai duyệt tri thức nguồn, ai xử lý khiếu nại, và các câu trả lời "ảnh hưởng quyền lợi" được gắn nhãn, được lưu nhật ký, và có cơ chế chuyển người thật khi độ chắc chắn thấp.

Anti definition: Không phải bài giảng đạo đức, cũng không phải lời cam kết treo tường, càng không phải "tuyên ngôn cho đẹp".

Lỗi hay gặp: Đặt KPI "đúng nhanh" mà không đặt KPI "đúng cách", ví dụ không đo tỷ lệ khiếu nại, tỷ lệ phải can thiệp thủ công, tỷ lệ câu trả lời bị sửa, hoặc tỷ lệ sự cố lộ thông tin.

Gợi ý kiểm chứng: Hỏi ba câu: "bằng chứng tin cậy nằm ở đâu", "ai có quyền dừng hoặc giảm tự động hóa", "người bị ảnh hưởng có đường khiếu nại và được xem xét lại không".

4.2. Ba lực đẩy làm rủi ro AI tăng nhanh

Định nghĩa: Rủi ro AI tăng mạnh khi một quyết định được nhân bản ở quy mô lớn, được tự động hóa, và dựa vào dữ liệu không hoàn hảo, từ đó lỗi nhỏ trở thành tác hại lớn.

Vì sao quan trọng: Cùng một sai lệch 1%, ở quy mô nhỏ là vài trường hợp, ở quy mô lớn là hàng nghìn người. Khi đã tự động hóa, con người rút khỏi điểm chốt, nên tốc độ lan truyền của sai còn nhanh hơn tốc độ sửa.

Ví dụ: Một mô hình lọc hồ sơ loại nhầm 2% ứng viên, khi tuyển 1.000 người là 20 người, khi tuyển 1.000.000 người là 20.000 người. Nếu hệ thống còn tự động gửi email từ chối, không có kênh rà soát, tổn thất niềm tin sẽ phình ra theo cấp số nhân.

Anti definition: Không phải "AI nguy hiểm vì AI", mà nguy hiểm vì cách AI được đặt vào vận hành.

Lỗi hay gặp: Triển khai thử nghiệm xong rồi nhân bản, nhưng không có giám sát drift, không có ngưỡng dừng, không có vòng phản hồi từ khiếu nại, và không có cơ chế giảm tự động hóa khi thấy dấu hiệu bất thường.

Gợi ý kiểm chứng: Kiểm tra nhanh 4 yếu tố, quy mô tác động, mức tự động hóa, độ nhạy dữ liệu, và độ khó giải thích cho người bị ảnh hưởng.

4.3. Tác hại không đồng nghĩa với sai số

Định nghĩa: Sai số (errors) là chênh lệch dự đoán so với "nhãn đúng" hoặc mục tiêu kỹ thuật, còn tác hại (harms) là hậu quả thực tế lên con người và hệ thống, thường đi qua các kênh như thiên lệch, xâm phạm riêng tư, thiếu minh bạch, mất an toàn, và đứt gãy trách nhiệm. Nói ngắn gọn, sai số nằm trong phòng lab, tác hại nằm ngoài đời.

Vì sao quan trọng: Mô hình "đúng kỹ thuật" vẫn có thể gây hại nếu mục tiêu tối ưu sai, hoặc nếu quy trình vận hành biến một sai số nhỏ thành quyết định tước quyền lợi, loại trừ cơ hội.

Ví dụ: Chatbot trả lời nghe có vẻ đúng câu chữ nhưng "bịa" chính sách hoàn tiền, gây tranh chấp hàng loạt. Đây là tác hại vận hành, không chỉ là lỗi câu trả lời.

Anti definition: Không đánh đồng "độ chính xác cao" với "đáng tin cậy". Cũng không đánh đồng "không có lỗi lớn" với "không có hại".

Lỗi hay gặp: Đo mỗi chỉ số đo của khoa học dữ liệu (data science), bỏ quên chỉ số đo của niềm tin, ví dụ tỷ lệ phàn nàn, tỷ lệ phải thực hiệncơ chế leo thang xử lý, tỷ lệ quyết định bị khiếu nại, tỷ lệ người dùng bỏ dịch vụ sau khi bị AI xử lý.

Gợi ý kiểm chứng: Bắt buộc điền năm nhóm tác hại trong mọi đề xuất tình huống sử dụng, trước khi bàn đến mô hình sử dụng hay cách đo lường.

4.4. Hai lớp rủi ro: kỹ thuật và quản trị xã hội

Định nghĩa: Rủi ro AI không chỉ đến từ sai số kỹ thuật của mô hình, mà còn đến từ cách hệ thống được thiết kế, triển khai, giám sát và sử dụng trong bối cảnh xã hội, hai lớp này tương tác và khuếch đại lẫn nhau.

Vì sao quan trọng: Nếu chỉ nhìn rủi ro như lỗi kỹ thuật, tổ chức sẽ tối ưu mô hình mà bỏ sót những tác hại phát sinh từ quyết định về tổ chức, phân quyền, quy trình, và quan hệ quyền lực, nơi hậu quả thường lớn hơn và khó đảo ngược hơn.

Mini story 3: SyRI - Sự sụp đổ của "hộp đen" giám sát

La bàn nội dung: Hai lớp rủi ro, kỹ thuật và quản trị xã hội

Khái niệm cốt lõi: Minh bạch (Transparency), Quyền riêng tư (Privacy)

Chính phủ Hà Lan từng phát triển một hệ thống có tên SyRI (System Risk Indication) để chống gian lận phúc lợi. SyRI được mô tả là một công cụ liên kết dữ liệu từ nhiều nguồn hành chính khác nhau, để tạo "chỉ báo rủi ro" phục vụ phát hiện gian lận trong các lĩnh vực như phúc lợi, trợ cấp và thuế.

Tuy nhiên nhiều người dân bị gắn cờ rủi ro không hề biết vì sao mình bị chọn, dữ liệu nào đã được dùng, hay logic của mô hình vận hành ra sao. Nó trở thành một "hộp đen" (black box), theo đúng nghĩa pháp lý, người chịu tác động không có cách kiểm chứng và tự bảo vệ.

Tháng 2/2020, một liên minh các tổ chức xã hội dân sự và công dân kiện Nhà nước. Tòa án Quận The Hague (The Hague District Court) ra phán quyết rằng khuôn khổ pháp lý cho SyRI vi phạm Điều 8 của Công ước Châu Âu về Nhân quyền (quyền riêng tư). Lập luận trọng tâm là, trong hình thức hiện tại, việc áp dụng SyRI "thiếu minh bạch và khả kiểm" (insufficiently transparent and verifiable), nên không đạt được "cân bằng hợp lý" (fair balance) giữa lợi ích cộng đồng và quyền riêng tư của cá nhân. Đáng chú ý, vụ kiện liên quan SyRI đặt trọng tâm vào câu hỏi, Nhà nước được phép đi xa đến đâu khi dùng công nghệ mới để "soi" đời sống người dân.

Vụ việc SyRI làm rõ "hai lớp rủi ro", một là lớp kỹ thuật, hệ thống được thiết kế để tạo chỉ báo rủi ro, hai là lớp quản trị xã hội, nơi tính hợp pháp, minh bạch, và quyền con người quyết định hệ thống có được chấp nhận hay không. Nó nhắc lại một nguyên tắc đạo đức cốt lõi, minh bạch (transparency) không chỉ là tính năng kỹ thuật, mà là điều kiện nền để xã hội còn có thể kiểm soát quyền lực mới mà AI tạo ra (Henley et al., 2020).

Tài liệu tham khảo

1. Henley, J., & Booth, R. (2020, February 5). Welfare surveillance system violates human rights, Dutch court rules. The Guardian. https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules
2. The Hague District Court. (2020, February 5). SyRI legislation in breach of European Convention on Human Rights. Rechtspraak. https://www.rechtspraak.nl/organisatie-en-contact/organisatie/rechtbanken/rechtbank-den-haag/nieuws/syri-legislation-in-breach-of-european-convention-on-human-rights

Anti definition: Không phải mọi rủi ro đều có thể giải quyết bằng cải thiện dữ liệu hay thuật toán, và cũng không phải cứ có chính sách là rủi ro kỹ thuật tự biến mất.

Gợi ý kiểm chứng: Khi một hệ thống gây tranh cãi, hãy tách câu hỏi thành hai lớp, lỗi nằm ở mô hình hay ở cách ra quyết định, phân quyền, giám sát và xử lý phản hồi.

4.5. Đứt gãy trách nhiệm và quyền phản đối

Định nghĩa: Đứt gãy trách nhiệm (accountability break) xảy ra khi quyết định được tự động hóa nhưng không rõ ai có thẩm quyền dừng, sửa, giải thích, và ghi vết. Quyền phản đối (contestability) là khả năng người bị ảnh hưởng có đường phản đối, và được xem xét lại bằng một quy trình rõ ràng, có thời hạn.

Vì sao quan trọng: Khi không có "cầu thang trách nhiệm", sự cố sẽ biến thành đổ lỗi, và tổ chức mất khả năng tự bảo vệ bằng bằng chứng. Tệ hơn, người dùng cảm thấy bị đối xử bởi một bức tường vô danh.

Ví dụ: Khách hàng bị từ chối tín dụng, không có kênh giải thích, không có cơ chế rà soát thủ công, không có nhật ký về yếu tố nào đóng góp lớn, và không có Thỏa thuận mức dịch vụ (Service Level Agreement, SLA) cho phản hồi khiếu nại.

Anti definition: Trách nhiệm giải trình không phải tìm người chịu tội, mà là thiết kế trách nhiệm và bằng chứng. Quyền phản đối không phải "ai cũng cãi được", mà là có quy trình để sửa sai hợp lệ.

Lỗi hay gặp: Giao hết cho nhà cung cấp, rồi phát hiện nhà cung cấp cũng không có nhật ký và ghi vết đủ sâu, hoặc không hỗ trợ cơ chế khiếu nại và rà soát (review) theo ngữ cảnh nghiệp vụ.

Gợi ý kiểm chứng: Có ma trận phân công trách nhiệm (RACI) rõ, có quy trình thực hiện cơ chế leo thang xử lý, có nhật ký quyết định, và có kênh tiếp nhận phản ánh với Thỏa thuận mức dịch vụ (SLA) cụ thể, ít nhất phải định nghĩa ai nhận, ai xử, ai phê, và trong bao lâu.

4.6. Ảnh hưởng và bối cảnh quyết định mức nghiêm ngặt

Định nghĩa: Không phải mọi tình huống sử dụng đều cần mức nghiêm ngặt như nhau, mức nghiêm ngặt phải tăng theo mức ảnh hưởng và rủi ro của bối cảnh.

Vì sao quan trọng: Vì nguồn lực có hạn, nếu đặt mức nghiêm ngặt sai chỗ, tổ chức sẽ trả giá bằng mất niềm tin, tranh chấp, và khủng hoảng.

Ví dụ: Những tình huống sử dụng phân bổ cơ hội và quyền lợi, hoặc liên quan an toàn và dữ liệu nhạy cảm, thường đòi hỏi mức nghiêm ngặt cao hơn.

Anti definition: Đây không phải tư duy cấm đoán, cũng không phải tư duy làm liều để lấy nhanh kết quả, lỗi hay gặp là chọn "dễ thắng" đúng vào nơi ảnh hưởng lớn, rồi chịu hậu quả bằng khủng hoảng niềm tin.

Gợi ý kiểm chứng: Hỏi bốn câu: hệ thống động đến quyền lợi không, có rủi ro an toàn không, dữ liệu có nhạy cảm không, mức tự động hóa đến đâu, phần "gắn mức kiểm soát cụ thể" sẽ được làm rõ ở các chương về quản trị và vận hành.

4.7. Sự cố như hệ quả của chuỗi nguyên nhân quản trị

Định nghĩa: Sự cố AI hiếm khi là một điểm lỗi đơn lẻ, mà thường là kết quả của một chuỗi quyết định quản trị, từ chọn bài toán, dữ liệu, mức tự động hóa, đến giám sát và phản ứng khi có cảnh báo.

Vì sao quan trọng: Cách nhìn này giúp tổ chức rời khỏi tư duy "đổ lỗi

cho mô hình", để nhìn thẳng vào các điểm chốt quản trị đã bị bỏ qua, nơi nếu sửa sớm có thể ngăn sự cố leo thang thành khủng hoảng.

Anti definition: Không phải mọi sự cố đều là tai nạn bất khả kháng, và cũng không phải cứ vá kỹ thuật sau sự cố là đã giải quyết tận gốc vấn đề.

Gợi ý kiểm chứng: Khi xảy ra sự cố, lần ngược lại các quyết định đã được đưa ra trước đó, ai quyết, dựa trên thông tin nào, thiếu cảnh báo gì, và phản hồi bị chậm ở khâu nào.

5. DANH SÁCH TỰ RÀ SOÁT

1. Tôi nói được trong 30 giây vì sao đạo đức AI là yêu cầu vận hành, không chỉ học thuật. Nếu chưa, viết một câu theo cấu trúc "AI đang tự động hóa quyết định X, nên phải kiểm soát Y". (LO1, LO9)
2. Tôi nêu được ba lực đẩy quy mô, tự động hóa, phụ thuộc dữ liệu bằng ví dụ thật của tổ chức. Nếu chưa, chọn một tình huống sử dụng và trả lời ba câu hỏi, quy mô, tự động hóa, dữ liệu. (LO2, LO9)
3. Tôi phân biệt được rủi ro kỹ thuật và rủi ro xã hội, pháp lý, quản trị bằng cùng một tình huống. Nếu chưa, liệt kê 5 tác hại rồi gắn nhãn "kỹ thuật" hay "quản trị". (LO3, LO10)
4. Tôi liệt kê được 5 nhóm tác hại và biết nhóm nào dễ xảy ra nhất với tổ chức mình. Nếu chưa, xem lại sự cố, khiếu nại, và quy trình vận hành gần nhất rồi gắn vào 5 nhóm. (LO4, LO11)
5. Tôi biết hỏi "ai bị ảnh hưởng, ảnh hưởng gì" trước khi hỏi "độ chính xác bao nhiêu". Nếu chưa, thêm ô "nhóm bị ảnh hưởng" vào mọi tài liệu đề xuất tình huống sử dụng. (LO5, LO12)
6. Tôi xếp được mức ảnh hưởng theo 3 lớp, con người, tổ chức, xã hội. Nếu chưa, dùng thang 3 mức và ghi lý do 1 câu cho mỗi lớp. (LO5, LO8)
7. Tôi nhận diện được dấu hiệu tình huống sử dụng rủi ro cao ngay từ đầu. Nếu chưa, kiểm tra 4 tiêu chí, quyền lợi, an toàn, dữ liệu nhạy cảm, mức tự động hóa. (LO7, LO2)
8. Tôi lập được shortlist top 5 tình huống sử dụng rủi ro cao của tổ

chức. Nếu chưa, bắt đầu từ 10 tình huống sử dụng hay nhất, rồi dùng impact tiering để lọc. (LO7, LO5)

9. Khi gặp sự cố AI, tôi không dừng ở "lỗi model". Nếu chưa, buộc nhóm viết một trang (one pager) chuỗi nguyên nhân quản trị trước khi đề xuất đổi model. (LO6, LO10)
10. Tôi chỉ ra được ít nhất 2 điểm đứt gãy trách nhiệm. Nếu chưa, vẽ RACI cho ai ra quyết định, ai giám sát, ai xử lý khiếu nại. (LO6, LO4)
11. Tôi nối được đạo đức AI với ít nhất 2 lợi ích kinh doanh cụ thể. Nếu chưa, viết 2 kịch bản, chi phí sự cố, chi phí kiểm soát sớm, và so sánh. (LO8, LO1)
12. Tôi hiểu đánh đổi (trade off) phổ biến, hiệu quả và an toàn, chính xác và công bằng, minh bạch và riêng tư. Nếu chưa, chọn 1 đánh đổi và viết điều kiện chấp nhận được, ai duyệt. (LO8, LO3)
13. Tôi có thông điệp ngắn, nhất quán để nói với lãnh đạo về vì sao cần ưu tiên đạo đức AI. Nếu chưa, luyện nói 3 câu, vấn đề, rủi ro, quyết định cần chốt. (LO9, LO11)
14. Tôi nhận diện được ít nhất 2 hiểu nhầm phổ biến và biết cách gỡ. Nếu chưa, làm bảng "myth, reality, bằng chứng", dùng trong họp nội bộ. (LO10, LO1)
15. Tôi phác thảo được "lập trường ban đầu" của dự án hoặc tổ chức trong 10 dòng. Nếu chưa, dùng khung "chúng tôi ưu tiên, chúng tôi không chấp nhận, chúng tôi yêu cầu bằng chứng". (LO11, LO9)
16. Tôi có bộ câu hỏi khởi động để dẫn sang chuẩn hóa thuật ngữ ở chương 2. Nếu chưa, liệt kê 10 thuật ngữ hay cãi nhất trong dự án và đánh dấu cần chuẩn hóa. (LO12, LO10)
17. Tôi có bộ câu hỏi khởi động để dẫn sang vòng đời hệ thống ở chương 4. Nếu chưa, vẽ vòng đời tình huống sử dụng và đánh dấu pha rủi ro cao nhất. (LO12, LO5)
18. Tôi có bộ câu hỏi khởi động để dẫn sang khung quản trị ở chương 5. Nếu chưa, xác định chủ dữ liệu (owner), người rà soát (reviewer),

người phê chuẩn (approver), và yêu cầu nhật ký và ghi vết, dấu vết kiểm toán. (LO12, LO7)

Mẹo thực hành

Một là chọn 1 tình huống sử dụng thực tế, thực hiện "Khung rủi ro, tác động, lợi ích" (Risk, impact, benefit canvas) trong 15 phút, đem vào họp tuần tới, coi như một bài kiểm tra sức khỏe, xem nhóm có nói được cùng một ngôn ngữ và chốt được điều kiện tối thiểu không.

Hai là đừng để "độ chính xác" che khuất "tác hại", luôn ghi 5 nhóm tác hại: thiên kiến, quyền riêng tư, minh bạch, an toàn, trách nhiệm giải trình, ở trang đầu tài liệu, trước cả phần chọn model và chọn dữ liệu.

Ba là khi tranh luận nội bộ, chốt "ai chịu trách nhiệm" trước khi chốt "model nào", vì mô hình có thể thay, nhưng trách nhiệm và cơ chế chịu trách nhiệm phải được thiết kế ngay từ đầu.

Bốn là nếu có dữ liệu cá nhân, ghi rõ căn cứ xử lý dữ liệu, mục đích, phạm vi, phân quyền, và nhật ký xử lý theo Luật Bảo vệ dữ liệu cá nhân 2025 (Luật Bảo vệ dữ liệu cá nhân, 2025) và hướng dẫn từ Nghị định 356/2025/NĐ-CP ngày 31/12/2025, đừng để đến lúc có sự cố mới lục lại xem mình dựa vào cái gì.

Năm là làm việc với nhà cung cấp, yêu cầu tối thiểu phải có nhật ký và ghi vết, dấu vết kiểm toán, cơ chế xử lý sự cố, và hỗ trợ kênh khiếu nại, không có thì không pilot, vì bạn không thể thuê ngoài trách nhiệm giải trình.

Câu hỏi suy ngẫm

Tình huống sử dụng AI nào của tổ chức tôi có thể gây tác hại lớn nhất? nếu nó sai 1%, và tác hại đó đi qua kênh nào trong 5 nhóm như thế nào?

Điểm đứt gãy trách nhiệm dễ xảy ra nhất trong tổ chức tôi đang nằm ở đâu? ở khâu mục tiêu, dữ liệu, phê duyệt, vận hành, hay phản hồi và truyền thông?

Nếu ngày mai có sự cố, tôi có "bằng chứng" gì để giải thích? và sửa nhật ký nào? biên bản nào, quy trình nào, hay cuối cùng chỉ có lời xin lỗi?

6. CẦU NỐI GIỮA CÁC CHƯƠNG VÀ ĐIỂM KIỂM 90 NGÀY

Chương 1 đặt nền tảng bằng một quan điểm rất thực tế: đạo đức AI là vấn đề vận hành và trách nhiệm, không phải yếu tố trang trí. Bạn đã có cách nhìn rủi ro theo hai lớp, kỹ thuật và quản trị xã hội, và có một công cụ một trang (one pager tool) để biến tranh luận thành quyết định (Risk, impact, benefit canvas). Nhưng để dùng công cụ này không bị "mỗi

người hiểu một kiểu", chương 2 sẽ chuẩn hóa ngôn ngữ chung với những thuật ngữ công bằng (fairness), minh bạch (transparency) và khả năng giải thích (explainability), trách nhiệm giải trình (accountability), quyền riêng tư (privacy), an toàn (safety) và an ninh (security), khả năng phản đối (contestability), để khi bạn nói "tác hại", "rủi ro cao", hay "điều kiện tối thiểu", cả nhóm hiểu giống nhau và đo được giống nhau.

Từ chương 2, ta sẽ đi tiếp sang chương 4 để đặt đạo đức đúng vị trí trong vòng đời hệ thống AI, từ mục tiêu, dữ liệu, huấn luyện, đánh giá, triển khai đến vận hành. Lý do là phần lớn "đứt gãy trách nhiệm" không xảy ra ở chỗ bạn chọn model gì, mà ở chỗ bạn đặt kiểm soát sai pha, ví dụ không kiểm định dữ liệu trước khi huấn luyện, không có giám sát drift sau triển khai, hoặc không có cơ chế phản hồi và khiếu nại ở tuyến đầu. Sau đó, chương 5 sẽ đưa bạn vào khung quản trị cụ thể, như NIST AI RMF, ISO/IEC 42001 và ISO/IEC 23894, để biến các câu hỏi trong canvas thành hệ thống vai trò, quy trình, bằng chứng, và cải tiến liên tục.

Điểm kiểm 90 ngày của chương 1 không phải là "đọc xong thấy hay", mà là có kết quả thật và thói quen vận hành mới. Trong 90 ngày, bạn nên hoàn thành ít nhất hai việc:

Việc thứ nhất, chọn tối thiểu 5 tình huống sử dụng AI đang có hoặc sắp làm, điền Risk, impact, benefit canvas, và buộc phải chốt quyết định triển khai, dừng hoặc triển khai có điều kiện, kèm điều kiện tối thiểu bằng ngôn ngữ kiểm chứng được, nhật ký và ghi vết gì, ai rà soát, cơ chế leo thang xử lý ra sao? Thỏa thuận mức dịch vụ khiếu nại thế nào?

Việc thứ hai, chọn 1 tình huống sử dụng rủi ro cao nhất trong 5 tình huống đã chọn, làm một buổi "mổ xẻ theo chuỗi nguyên nhân quản trị" như lab của chương, và tạo ra một danh sách hành động khắc phục, ai làm, làm trong bao lâu, bằng chứng gì để chứng minh đã làm.

Nếu bạn làm được hai việc này trong 90 ngày, bạn sẽ bước sang chương 2, 4, 5 với một lợi thế rất lớn, bạn không học thuật ngữ để nhớ, bạn học thuật ngữ để dùng, không học vòng đời để vẽ, bạn học vòng đời để đặt kiểm soát, và không học khung quản trị để trang trí, bạn học để tổ chức có thể tự chứng minh mình đáng tin cậy.

CÂU CHỐT CHƯƠNG

"Đạo đức AI là quy trình và bằng chứng, không phải lời hứa."

Tài liệu tham khảo

1. Bansal, D. (2025). Embedding Ethics in the Heart of Digital Revolution. In Artificial Intelligence, Entrepreneurship and Risk (pp. 85–100). Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-45544-6_5
2. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. https://fairmlbook.org/
3. Dobbe, R. I. J. (2022). System Safety and Artificial Intelligence. In The Oxford Handbook of AI Governance (pp. 441–458). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.67
4. Erk, C. (2025). Normative Management: Why Management Cannot Do Without Philosophy. In Artificial Intelligence, Entrepreneurship and Risk (pp. 59–84). Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-45544-6_4
5. Floridi, L. (2023). The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001
6. Greene, K. G. (2022). AI Governance Multi-stakeholder Convening. In The Oxford Handbook of AI Governance (pp. 109–126). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.6
7. Hudson, V. M. (2022). Standing Up a Regulatory Ecosystem for Governing AI Decision-making. In The Oxford Handbook of AI Governance (pp. 276–298). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.17
8. Luật Bảo vệ dữ liệu cá nhân, Luật số 91/2025/QH15 (2025).
9. National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1; p. NIST AI 100-1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1
10. National Institute of Standards and Technology (NIST). (2024). Artificial intelligence risk

management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.600-1

11. Nghị định 13/2023/NĐ-CP (2023). https://vanban.chinhphu.vn/?pageid=27160&docid=210487
12. Nghị định 356/2025/NĐ-CP, 356/2025/NĐ-CP (2025). https://vanban.chinhphu.vn/?pageid=27160&docid=216387
13. Prem, E. (2025). Managing Entrepreneurial AI Ethics Risks. In Artificial Intelligence, Entrepreneurship and Risk (pp. 453–475). Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-45544-6_23
14. Raji, I. D. (2023). The Anatomy of AI Audits. In The Oxford Handbook of AI Governance (pp. 495–516). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.28
15. Regulation (EU) 2024/1689 on Artificial Intelligence, OJ L § 2024/1689 (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401689
16. Smuha, N. A., & Yeung, K. (2025). The European Union's AI Act: Beyond Motherhood and Apple Pie? In N. A. Smuha (Ed.), The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence. Cambridge University Press.

THỰC HÀNH

CÔNG CỤ KHUNG VÀ SẢN PHẨM ĐẦU RA

Công cụ chính: Khung rủi ro, tác động, lợi ích (Risk, impact, benefit canvas)

Mục tiêu: Biến tranh luận đạo đức thành quyết định quản trị và ưu tiên kiểm soát, theo tư duy quản trị dựa trên rủi ro làm ngay ở giai đoạn chọn tình huống sử dụng và buổi khởi động. (National Institute of Standards and Technology (NIST), 2023)

Khung (canvas) 1 trang, các ô điền

1. Mô tả tình huống sử dụng, 5 đến 7 dòng

AI làm gì, đưa ra đề xuất hay quyết định, tác động đến quy trình nào, cho ai, đầu ra dùng để làm gì.

2. Nhóm người bị ảnh hưởng

Trực tiếp, gián tiếp, nội bộ, bên ngoài, ai "chịu quyết định", ai "chịu hậu quả", ai "phải giải thích".

3. Dữ liệu chính và kiểm tra dữ liệu cá nhân

- Có dữ liệu cá nhân không, có dữ liệu cá nhân nhạy cảm không.
- Căn cứ xử lý dữ liệu, mục đích, phạm vi, thời hạn lưu.
- Phân quyền truy cập, nguyên tắc tối thiểu, nhật ký xử lý, ai được xem, ai được xuất.
- Ghi chú tuân thủ theo Nghị định 13/2023/NĐ-CP, đặc biệt với nghĩa vụ bảo vệ dữ liệu cá nhân và dữ liệu nhạy cảm.

4. Mức tự động hóa (automation level)

AI chỉ gợi ý, hay ra quyết định, hay tự động thực thi. Có người duyệt không, duyệt ở điểm nào, có quyền dừng, quyền hạ mức tự động hóa không (Regulation (EU) 2024/1689 on Artificial Intelligence, 2024).

5. Nhóm tác hại có thể xảy ra

Tick nhanh và ghi 1 đến 2 câu cho mỗi nhóm:

- Thiên kiến
- Quyền riêng tư
- Minh bạch và khả năng giải thích (transparency and explainability)
- An toàn và an ninh
- Trách nhiệm giải trình

6. Mức ảnh hưởng (impact tier), thấp, vừa, cao, theo 3 lớp

- Con người, ảnh hưởng quyền lợi, cơ hội, an toàn, phẩm giá
- Tổ chức, ảnh hưởng vận hành, tài chính, uy tín, tuân thủ
- Xã hội, ảnh hưởng niềm tin, công bằng, dịch vụ thiết yếu
  Ghi rõ "điểm đau lớn nhất" ở mỗi lớp, nếu có.

7. Lợi ích kinh doanh kỳ vọng (benefits)

ROI dự kiến, tiết kiệm thời gian, tăng chất lượng, giảm sai sót, tăng trải nghiệm, tăng uy tín, mở rộng thị trường. Ghi rõ thước đo, ví dụ giảm thời gian xử lý bao nhiêu phần trăm, giảm khiếu nại bao nhiêu.

8. Mức ưu tiên kiểm soát và quyết định triển khai

Chọn một trong ba: **Triển khai, dừng hoặc triển khai có điều kiện (Go, No go, Go with conditions)**.

Nếu triển khai có điều kiện (go with conditions)", liệt kê tối đa 3 điều kiện tối thiểu, ví dụ bắt buộc người phải rà soát (human review) ở bước X, bắt buộc nhật ký và ghi vết Y, bắt buộc kênh khiếu nại với SLA Z, bắt

buộc kiểm thử công bằng theo nhóm A. (National Institute of Standards and Technology (NIST), 2023)

9. Câu hỏi cần trả lời ở chương 2, 4, 5

- Chương 2, thuật ngữ nào phải chốt để khỏi nói lệch, công bằng, minh bạch, trách nhiệm giải trình, quyền riêng tư, an toàn, khả năng phản đối.
- Chương 4, pha vòng đời nào rủi ro nhất, dữ liệu, huấn luyện, đánh giá, triển khai, vận hành.
- Chương 5, khung quản trị nào cần dùng, NIST AI RMF cho quản trị rủi ro, ISO/IEC 42001 cho hệ thống quản lý AI, ISO/IEC 23894 cho hướng dẫn quản trị rủi ro AI.

Cách dùng trong 10 đến 15 phút

- Dùng canvas này ngay trong buổi khởi động, hoặc cuộc họp chọn tình huống sử dụng, đừng chờ đến khi đã code xong.
- Điền nhanh theo nhóm liên chức năng, tối thiểu phải có sản phẩm, dữ liệu, pháp chế, vận hành, nếu có thể thêm an ninh thông tin.
- Chốt 3 điều trước khi rời phòng, **tác hại lớn nhất**, **điểm đứt gãy trách nhiệm dễ nhất**, **điều kiện tối thiểu để được phép chạy thử nghiệm**. Nếu không chốt được, mặc định là "triển khai có điều kiện" ("Go with conditions") hoặc "Dừng" ("No go").

Sản phẩm đầu ra

Các sản phẩm đầu ra (Deliverables) có thể xây dựng bao gồm:

- **Sản phẩm 1:** phê duyệt bước đầu, map LO5, LO7, LO8, LO9.
- **Sản phẩm 2:** 01 quyết định ưu tiên kiểm soát, Triển khai, dừng hoặc triển khai có điều kiện (Go, No go, Go with conditions), kèm tối đa 3 điều kiện, map LO7, LO10, LO11.
- **Sản phẩm 3:** 01 danh sách câu hỏi chuyển sang chương 2, 4, 5, map LO9, LO10, LO12.

Góc chuyên gia

Tôi gặp nhiều dự án AI thất bại không phải vì mô hình dở, mà vì họ không có một tờ giấy để thống nhất ai chịu trách nhiệm và ở chỗ nào. Người ta tranh luận về đạo đức bằng cảm xúc, rồi cuối cùng quay về KPI, làm cho xong. Canvas này là cách tôi ép cuộc họp đi vào quyết định, rủi ro gì, ai bị ảnh hưởng, dữ liệu có sạch và có căn cứ không, ai được quyền dừng, và muốn chạy thử nghiệm thì phải có điều kiện tối thiểu nào. Nó không làm bạn "đạo đức hơn", nó làm bạn "đỡ sai sót hơn" trong vận hành, tôi muốn bạn nhớ điều này.

Tôi từng thấy nhiều tổ chức lao vào AI như lao vào một cuộc đua, ai cũng thích tốc độ, ít ai chịu nhìn phanh. Nhưng đời giống đường đèo, sương mù dày, nếu không có phanh thì nhanh chỉ là cách để đến tai nạn sớm hơn. Đạo đức AI, với tôi, không phải làm cho bạn "tốt hơn" về đạo lý mà làm cho bạn "đỡ ngây thơ hơn" trước những quyết định có tác động lớn. Bạn càng muốn đi nhanh, bạn càng phải nói rõ ai cầm lái, ai đạp phanh, ai ghi vết, ai lắng nghe khiếu nại, tôi muốn bạn nhớ điều này.

BÀI TẬP THỰC HÀNH

Chẩn đoán một sự cố AI, từ "lỗi model" sang "lỗi quản trị"

Bối cảnh, tình huống sử dụng và dữ liệu mẫu

1. Tình huống sử dụng: chatbot tư vấn khoản vay và thu hồi nợ.
2. Quyết định liên quan: gợi ý hạn mức, gợi ý nhắc nợ, phân loại khách hàng cần gọi điện.
3. Dữ liệu dùng: lịch sử giao dịch, lịch sử chat, thông tin định danh cơ bản, nhật ký và ghi vết cuộc gọi.
4. Sự cố giả lập, 3 hiện tượng:

- Chatbot nhắc sai chính sách phạt.
- Hội thoại có lẫn dữ liệu cá nhân (personal data) của khách hàng khác.
- Nhóm lao động tự do bị phân loại "rủi ro cao" bất thường.

Dữ liệu mẫu, trích 5 bản ghi

1. C001, nhóm nghề, lao động tự do, gợi ý hạn mức 15 triệu, nhắc phạt 5% sai chính sách, risk score 0.82.

2. C002, nhóm nghề, văn phòng, gợi ý hạn mức 60 triệu, trả lời đúng chính sách, risk score 0.41.
3. C003, nhóm nghề, lao động tự do, gợi ý hạn mức 12 triệu, bị xếp cần gọi điện, risk score 0.79.
4. C004, nhóm nghề, tiểu thương, chatbot trả lẫn "Số CCCD của C002", risk score 0.52.
5. C005, nhóm nghề, văn phòng, nhắc nợ đúng nhưng không giải thích vì sao bị gọi điện, risk score 0.63.

Các bước

1. Liệt kê tác hại theo 5 nhóm: thiên kiến, minh bạch, trách nhiệm giải trình, quyền riêng tư, an toàn.
2. Tách rủi ro kỹ thuật và rủi ro xã hội, pháp lý, quản trị.
3. Vẽ chuỗi nguyên nhân quản trị: mục tiêu tối ưu, dữ liệu, điểm chốt con người, giám sát, phản hồi, trách nhiệm.
4. Điền Risk, impact, benefit canvas và chốt: Triển khai, dừng hoặc triển khai có điều kiện.
5. Viết 10 dòng "điều kiện tối thiểu" để chạy thử nghiệm, gồm nhật ký và ghi vết, cơ chế leo thang xử lý, xử lý khiếu nại, phân quyền dữ liệu.

Đầu ra

1. 1 trang phân tích chuỗi nguyên nhân quản trị.
2. 1 rủi ro (risk), tác động (impact), lợi ích (benefit) canvas hoàn chỉnh.

Thang chấm điểm (Rubric)

1. Phân biệt đúng rủi ro kỹ thuật và rủi ro quản trị, tối thiểu 3 ý mỗi nhóm.
2. Xác định đúng điểm đứt gãy trách nhiệm, đề xuất cơ chế khả năng khiếu nại cụ thể.

3. Đánh giá mức ảnh hưởng có căn cứ, quyết định triển khai, dừng hoặc triển khai có điều kiện hợp lý.
4. Đề xuất ít nhất 2 kiểm soát cụ thể, policy, nhật ký và ghi vết, rà soát, cơ chế leo thang xử lý, gắn vào đúng điểm trong quy trình.



Góc chuyên gia

Trong binh pháp, người ta nói thắng thua nằm ở chuẩn bị, nhưng trong AI, chuẩn bị nằm ở câu hỏi. Bạn hỏi sai, bạn chạy nhanh đến sai. Tôi hay thấy người ta cãi nhau về model nào mạnh, lời nhắc nào hay, nhưng lại né câu hỏi khó, ai bị ảnh hưởng, ai chịu trách nhiệm, dữ liệu có hợp pháp không, có đường khiếu nại không. Khi bạn tập thói quen đặt câu hỏi đúng, bạn đang xây một đức hạnh hiện đại, sự cẩn trọng. Nó không kìm chân đổi mới, nó cứu bạn khỏi những cú trả giá nguội lạnh, tôi muốn bạn nhớ điều này.