TCVN **TIÊU CHUẨN QUỐC GIA**

TCVN 13903:2023
ISO/IEC TR 24028:2020

Xuất bản lần 1

CÔNG NGHỆ THÔNG TIN – TRÍ TUỆ NHÂN TẠO – TỔNG QUAN VỀ TÍNH ĐÁNG TIN CẬY TRONG TRÍ TUỆ NHÂN TẠO

Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

HÀ NỘI – 2023

Mục lục

Lời nói đầu

TCVN 13903:2023 hoàn toàn tương đương với ISO/IEC TR 24028:2020.

TCVN 13903:2023 do Viện Công nghiệp Phần mềm và Nội dung số Việt Nam biên soạn, Bộ Thông tin và Truyền thông đề nghị, Tổng cục Tiêu chuẩn Đo lường Chất lượng thẩm định, Bộ Khoa học và Công nghệ công bố.

Công nghệ thông tin – Trí tuệ nhân tạo – Tổng quan về tính đáng tin cậy trong trí tuệ nhân tạo

Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

1 Phạm vi áp dụng

Tiêu chuẩn này xem xét các vấn đề dưới đây liên quan đến tính đáng tin cậy trong các hệ thống AI:

- Các phương pháp tiếp cận để tạo lập sự tin cậy vào các hệ thống AI thông qua tính minh bạch, tính diễn giải, khả năng điều khiển v.v..;
- Các bẫy kỹ thuật và các mối đe dọa và rủi ro điển hình liên quan đến các hệ thống AI, các kỹ thuật và phương pháp giảm thiểu có thể; và
- Các phương pháp tiếp cận để đánh giá tính khả dụng, khả năng phục hồi, tính tin cậy, độ chính xác, an toàn, bảo mật và quyền riêng tư của các hệ thống AI.

Đặc tả về các mức độ đáng tin cậy đối với các hệ thống AI nằm ngoài phạm vi của tiêu chuẩn này.

2 Tài liệu viện dẫn

Tiêu chuẩn này không có tài liệu viện dẫn.

3 Thuật ngữ và định nghĩa

Tiêu chuẩn này sử dụng các thuật ngữ và định nghĩa dưới đây.

ISO và IEC duy trì cơ sở dữ liệu thuật ngữ sử dụng trong hoạt động tiêu chuẩn hóa tại các địa chỉ dưới đây:

- Nền tảng trình duyệt trực tuyến của ISO: tại địa chỉ https://www.iso.org/obp
- Từ vựng kỹ thuật điện của IEC: tại địa chỉ https://www.electropedia.org/

3.1

Trách nhiệm giải trình (accountability)

Thuộc tính đảm bảo rằng các hành động của một *thực thể* (3.16) có thể được truy nguyên duy nhất cho thực thể đó.

[nguồn: ISO/IEC 2382: 2015, 2126250 được sửa đổi – Xóa các chú thích].

3.2

Tác nhân (actor)

Thực thể (3.16) giao tiếp và tương tác.

[nguồn: ISO/IEC TR 22417:2017, 3.1].

3.3

Thuật toán (algorithm)

Tập hợp các quy tắc để biến đổi biểu diễn logic của dữ liệu (3.11).

[nguồn: ISO/IEC 11557:1992, 4.3].

3.4

Trí tuệ nhân tạo (artificial intelligence)

AI

Năng lực của một *hệ thống được thiết kế* (3.38) để thu nhận, xử lý và áp dụng các tri thức và kỹ năng.

CHÚ THÍCH 1: Tri thức là các dữ kiện, *thông tin* (3.20) và kỹ năng có được thông qua kinh nghiệm hoặc giáo dục.

3.5

Tài sản (asset)

Bất kỳ thứ gì có *giá trị* (3.46) đối với một *bên liên quan* (3,37).

CHÚ THÍCH 1: Có nhiều loại hình tài sản, bao gồm:

a) *Thông tin* (3.20);

b) phần mềm, chẳng hạn như chương trình máy tính;

c) Vật chất, chẳng hạn như máy tính;

d) Dịch vụ;

e) Con người và kỹ năng, trình độ, kinh nghiệm của họ; và

f) Vô hình, chẳng hạn như danh tiếng và hình ảnh.

[Nguồn: ISO/IEC 21827: 2008, 3.4 được sửa đổi - Trong định nghĩa, "tổ chức" đã được thay thế bằng "một bên liên quan". Xóa chú thích 1].

3.6

Thuộc tính (attribute)

Tích chất hoặc đặc điểm của một đối tượng có thể phân biệt về mặt định lượng hoặc định tính bằng con người hoặc các phương tiện tự động.

[nguồn: ISO/IEC/IEEE 15939:2017, 3.2].

3.7

Tự trị (autonomy)

Tính tự trị (autonomous)

Đặc điểm của một *hệ thống* (3.38) được quản trị bởi các quy tắc riêng của nó có được từ kết quả của quá trình tự học.

CHÚ THÍCH 1: Các hệ thống như vậy không chịu sự *điều khiển* (3.10) hoặc giám sát bên ngoài.

3.8

Thiên vị (bias)

Thiên vị đối với cái gì, người hoặc nhóm người nào hơn.

3.9

Tính nhất quán (consistency)

Mức độ đồng nhất, tiêu chuẩn hóa và không có mâu thuẫn giữa các tài liệu, thành phần hoặc các bộ phận của một *hệ thống* (3.38).

[nguồn: ISO/IEC 21827: 2008, 3.14].

3.10

Điều khiển (control)

Hành động có mục đích trên hoặc trong một *quá trình* (3.29) để đáp ứng các mục tiêu cụ thể.

[nguồn: IEC 61800-7-1: 2015, 3.2.6].

3.11

Dữ liệu (data)

Sự biểu diễn có thể biên dịch lại *thông tin* (3.20) theo cách thức chính quy, phù hợp để truyền thông, diễn giải, hoặc xử lý.

CHÚ THÍCH 1: *Dữ liệu* (3.11) có thể được xử lý bằng con người hoặc phương tiện tự động.

[nguồn: ISO/IEC 2382: 2015, 2121272 được sửa đổi – xóa bỏ chú thích 2 và 3].

3.12

Chủ thể dữ liệu (data subject)

Sự riêng biệt về *dữ liệu cá nhân* (3.27) được ghi nhận.

[nguồn: ISO 5127: 2017, 3.13.4.01 được sửa đổi – Bỏ chú thích 1].

3.13

Cây quyết định (decision tree)

Mô hình học tập có giám sát mà sự suy diễn trong đó được biểu diễn bằng cách duyệt qua một hoặc nhiều cấu trúc dạng cây.

3.14

Tính hiệu quả (effective)

Mức độ thực hiện các hoạt động theo kế hoạch và kết quả đạt được theo kế hoạch.

[nguồn: ISO 9000: 2015, 3.7.11 được sửa đổi – Bỏ chú thích 1].

3.15

Hiệu suất (efficiency)

Mối quan hệ giữa kết quả đạt được và nguồn lực được sử dụng.

[nguồn: ISO 9000: 2015, 3.7.10].

3.16

Thực thể (entity)

Bất kỳ điều gì cụ thể hoặc trừu tượng được quan tâm.

[nguồn: ISO/IEC 10746-2: 2009, 6.1].

3.17

Tổn hại (harm)

Thương tích hoặc thiệt hại cho sức khoẻ của con người hoặc thiệt hại về tài sản hoặc môi trường.

[nguồn: ISO/IEC Guide 51: 2014, 3.1].

3.18

Nguy cơ (hazard)

Nguồn tiềm ẩn gây ra sự *tổn hại* (3.17).

[nguồn: ISO/IEC Guide 51: 2014, 3.2].

3.19

Các yếu tố con người (human factors)

Các yếu tố môi trường, tổ chức và công việc kết hợp với các đặc điểm nhận thức con người ảnh hưởng đến hành vi của cá nhân hoặc tổ chức.

3.20

Thông tin (information)

Dữ liệu (3.11) có ý nghĩa.

[nguồn: ISO 9000: 2015, 3.8.2].

3.21

Tính toàn vẹn (integrity)

Thuộc tính bảo vệ sự chính xác và tính đầy đủ của *tài sản* (3.5).

[nguồn: ISO/IEC 27000: 2018, 3.36 được sửa đổi – Bổ sung định nghĩa "bảo vệ" trước "sự chính xác" và "của tài sản" đã được thêm vào sau "tính đầy đủ"].

3.22

Sử dụng theo dự kiến (intended use)

Sử dụng đúng với *thông tin* (3.20) được cung cấp về một sản phẩm hoặc *hệ thống* (3.38), hoặc trong trường hợp không có thông tin đó thì sử dụng theo *kiểu mẫu* (3.26) đã biết theo thông lệ.

[nguồn: ISO/IEC Guide 51: 2014, 3.6].

3.23

Học máy (machine learning)

ML

Quá trình (3.29) trong đó một khối chức năng cải thiện hiệu năng của nó bằng cách thu nhận tri thức hoặc kỹ năng mới; hoặc củng cố lại tri thức hoặc kỹ năng hiện có.

[nguồn: ISO/IEC 2382: 2015, 2123789].

3.24

Mô hình học máy (machine learning model)

Cấu trúc toán học tạo ra một suy diễn hoặc dự đoán dựa trên *dữ liệu* (3.11) hoặc thông tin đầu vào.

3.25

Mạng nơ-ron (neural network)

Mô hình điện toán xử lý phân tán, song song cục bộ và chứa một mạng các phần tử xử lý đơn giản được gọi là các nơ-ron nhân tạo thể hiện hành vi toàn cục phức tạp.

[nguồn: ISO 18115-1: 2013, 8.1].

3.26

Kiểu mẫu (pattern)

Tập hợp các thuộc tính và mối quan hệ của chúng được sử dụng để nhận dạng một *thực thể* (3.16) trong một ngữ cảnh nhất định.

[nguồn: ISO/IEC 2382: 2015, 2123798].

3.27

Dữ liệu cá nhân (personal data)

Dữ liệu (3.11) liên quan đến nhận dạng hoặc có thể nhận dạng được một cá nhân.

[nguồn: ISO 5127: 2017, 3.1.10.14 được sửa đổi – Bỏ các chú thích 1 và 2].

3.28

Sự riêng tư (privacy)

Tự do xâm nhập vào cuộc sống riêng tư hoặc vấn đề riêng của một cá nhân khi kết quả của sự xâm nhập đó là thu thập và sử dụng *dữ liệu* (3.11) quá mức hoặc bất hợp pháp về cá nhân đó.

[nguồn: ISO/IEC 2382: 2015, 2126263 được sửa đổi – Bỏ chú thích 1 và 2].

3.29

Quá trình (process)

Tập hợp các hoạt động có liên quan hoặc tương tác với nhau sử dụng các yếu tố đầu vào để mang lại kết quả dự kiến.

[nguồn: ISO 9000: 2015, 3.4.1 được sửa đổi – Bỏ các ghi chú].

3.30

Tính tin cậy (reliability)

Thuộc tính về sự phù hợp của hành vi và kết quả dự kiến.

[nguồn: ISO/IEC 27000: 2018, 3.55].

3.31

Rủi ro (risk)

Ảnh hưởng của tính bất định đến các mục tiêu.

CHÚ THÍCH 1: Ảnh hưởng là độ lệch so với cái được kỳ vọng. Nó có thể tích cực, tiêu cực hoặc cả hai, và nó tạo ra hoặc dẫn đến các cơ hội và mối đe dọa.

CHÚ THÍCH 2: Mục tiêu có thể có các khía cạnh và phạm trù khác nhau và có thể được áp dụng ở các cấp độ khác nhau.

CHÚ THÍCH 3: Rủi ro thường được thể hiện dưới dạng các nguồn gốc rủi ro, các sự kiện tiềm ẩn, hậu quả và khả năng xảy ra của chúng.

[nguồn: TCVN ISO 31000:2018, 3.1].

3.32

Người máy (robot)

Cơ chế dẫn động được lập trình với một mức độ *tự trị* (3.7), di chuyển trong môi trường của nó để thực hiện các tác vụ dự kiến.

CHÚ THÍCH 1: Người máy bao gồm hệ thống *điều khiển* (3.10) và giao diện của *hệ thống* (3.38) điều khiển.

CHÚ THÍCH 2: Việc phân loại người máy thành người máy công nghiệp hoặc người máy dịch vụ được thực hiện tùy theo ứng dụng dự kiến của nó.

[nguồn: ISO 18646-2: 2019, 3.1].

3.33

Khoa học người máy (robotics)

Khoa học và thực tiễn thiết kế, sản xuất và ứng dụng *người máy* (3.32).

[nguồn: ISO 8373: 2012, 2.16].

3.34

Tính an toàn (safety)

Không có *rủi ro* (3,31) không thể chấp nhận được.

[nguồn: ISO/IEC Chỉ dẫn 51: 2014, 3.14].

3.35

Tính bảo mật (security)

Mức độ mà sản phẩm hoặc *hệ thống* (3.38) bảo vệ *thông tin* (3.20) và *dữ liệu* (3.11) để mọi người hoặc các sản phẩm, hệ thống khác có mức độ truy cập dữ liệu phù hợp với các loại hình và cấp độ cho phép.

[nguồn: ISO/IEC 25010: 2011, 4.2.6].

3.36

Dữ liệu nhạy cảm (sensitive data)

Dữ liệu (3.11) với các tác động có hại tiềm tàng trong trường hợp bị tiết lộ hoặc sử dụng sai mục đích.

[nguồn: ISO 5127: 2017, 3.1.10.16].

3.37

Bên liên quan (stakeholder)

Bất kỳ cá nhân, nhóm hoặc tổ chức nào có thể ảnh hưởng, bị ảnh hưởng hoặc tự nhận thức bị ảnh hưởng bởi một quyết định hoặc hành động.

[NGUỒN: ISO/IEC 38500: 2015, 2.24].

3.38

Hệ thống (system)

Tổ hợp các phần tử tương tác được tổ chức để đạt được một hoặc nhiều mục đích đã định.

CHÚ THÍCH 1: Hệ thống đôi khi được coi như một sản phẩm hoặc các dịch vụ mà nó cung cấp.

[nguồn: ISO/IEC/IEEE 15288:2015, 3.38].

3.39

Mối đe dọa (threat)

Nguyên nhân tiềm ẩn của sự cố không mong muốn, có thể dẫn đến *tổn hại* (3.17) cho *hệ thống* (3.38), tổ chức hoặc cá nhân.

3.40

Huấn luyện (training)

Quá trình (3.29) thiết lập hoặc cải thiện các tham số của *mô hình học máy* (3.24) dựa trên *thuật toán* (3.3) học máy bằng cách sử dụng *dữ liệu* (3.11) huấn luyện.

3.41

Sự tin cậy (trust)

Mức độ *người dùng* (3.43) hoặc *bên liên quan* (3.37) khác tin tưởng rằng một sản phẩm hoặc *hệ thống* (3.38) sẽ hoạt động như dự kiến.

[nguồn: ISO/IEC 25010: 2011, 4.1.3.2].

3.42

Tính đáng tin cậy (trustworthiness)

Khả năng đáp ứng kỳ vọng của các *bên liên quan* (3.5.13) theo cách thức có thể xác minh được.

CHÚ THÍCH 1: Tùy thuộc vào bối cảnh hoặc lĩnh vực, cũng như sản phẩm hoặc dịch vụ cụ thể, dữ liệu và công nghệ được sử dụng, các đặc điểm khác nhau được áp dụng và cần được xác minh để đảm bảo đáp ứng kỳ vọng của các *bên liên quan* (3.5.13).

CHÚ THÍCH 2: Các đặc điểm của tính đáng tin cậy bao gồm: ví dụ như độ tin cậy, tính khả dụng, khả năng phục hồi, bảo mật, quyền riêng tư, an toàn, trách nhiệm giải trình, tính minh bạch, tính toàn vẹn, tính xác thực, chất lượng và khả năng sử dụng.

CHÚ THÍCH 3: Tính đáng tin cậy là một thuộc tính có thể được áp dụng cho các dịch vụ, sản phẩm, công nghệ, dữ liệu và thông tin, cũng như trong bối cảnh quản trị đối với các tổ chức.

3.43

Người dùng (user)

Cá nhân hoặc nhóm tương tác với *hệ thống* (3.38) hoặc hưởng lợi từ hệ thống trong quá trình sử dụng nó.

[nguồn: ISO/IEC/IEEE 15288:2015, 4.1.52 được sửa đổi – Bỏ chú thích 1].

3.44

Thẩm định (validation)

Xác nhận, thông qua việc cung cấp bằng chứng khách quan, rằng các yêu cầu sử dụng theo dự kiến hoặc ứng dụng cụ thể đã được đáp ứng.

CHÚ THÍCH 1: *Hệ thống* (3.38) đúng đắn đã được xây dựng.

[nguồn: ISO/IEC TR 29110-1: 2016, 3.73 được sửa đổi - Chỉ có câu cuối cùng của chú thích 1 được giữ lại và bỏ chú thích 2].

3.45

Giá trị (value)

<dữ liệu > đơn vị của *dữ liệu* (3.11).

[nguồn: ISO/IEC/IEEE 15939:2017, 3.41].

3.46

Giá trị (value)

<cộng đồng> (các) niềm tin mà một tổ chức dựa vào và các chuẩn mực mà tổ chức đó tìm kiếm để tuân thủ.

[nguồn: ISO 10303-11:2004, 3.3.22].

3.47

Xác minh (verification)

Xác nhận, thông qua việc cung cấp bằng chứng khách quan, rằng các yêu cầu cụ thể đã được đáp ứng.

CHÚ THÍCH 1: *Hệ thống* (3.38) đã được xây dựng đúng.

[nguồn: ISO/IEC TR 29110-1: 2016, 3.74 được sửa đổi - Chỉ có câu cuối cùng của chú thích 1 được giữ lại].

3.48

Tính dễ bị tổn thương (vulnerability)

Điểm yếu của một *tài sản* (3.5) hoặc *điều khiển* (3.10) có thể bị lợi dụng bởi một hoặc nhiều *mối đe dọa* (3.39).

[nguồn: ISO/IEC 27000: 2018, 3.77].

3.49

Khối lượng công việc (workload)

Hỗn hợp các tác vụ thường chạy trên một *hệ thống* (3.38) máy tính nhất định.

[nguồn: ISO/IEC/IEEE 24765:2017, 3.4618 được sửa đổi – Bỏ chú thích 1].

4 Tổng quan

Tiêu chuẩn này cung cấp một cái nhìn tổng quan về các chủ đề liên quan đến việc xây dựng tính đáng tin cậy của các hệ thống AI. Một trong những mục tiêu của tiêu chuẩn này này là hỗ trợ cộng đồng tiêu chuẩn xác định các lỗ hổng tiêu chuẩn hóa cụ thể trong lĩnh vực AI.

Điều 5 của tiêu chuẩn khảo sát sơ bộ các phương pháp tiếp cận hiện đang được sử dụng để xây dựng tính đáng tin cậy trong các hệ thống kỹ thuật và thảo luận về tiềm năng áp dụng của chúng đối với các hệ thống AI. Điều 6 của tiêu chuẩn xác định các bên liên quan. Điều 7 thảo luận về những mối quan tâm

của họ liên quan đến trách nhiệm, trách nhiệm giải trình, quản trị và tính an toàn của các hệ thống AI. Điều 8 của tiêu chuẩn khảo sát tính dễ bị tổn thương của hệ thống AI có thể làm giảm tính đáng tin cậy của chúng. Điều 9 của tiêu chuẩn xác định các biện pháp có thể để cải thiện tính đáng tin cậy của hệ thống AI bằng cách giảm thiểu tính dễ tổn thương trong vòng đời của nó. Các biện pháp bao gồm những khía cạnh liên quan đến cải thiện hệ thống AI về tính minh bạch, khả năng kiểm soát, xử lý dữ liệu, độ bền vững, thử nghiệm, đánh giá và sử dụng. Các kết luận được trình bày trong Điều 10.

5 Các khuôn khổ hiện có áp dụng cho tính đáng tin cậy

5.1 Cơ sở

Mục đích quan trọng của tiêu chuẩn này là cung cấp các định nghĩa chấp nhận được về hệ thống trí tuệ nhân tạo (AI) và tính đáng tin cậy.

Tiêu chuẩn này coi hệ thống AI là bất kỳ hệ thống nào (cho dù là sản phẩm hay dịch vụ) sử dụng AI. Có nhiều loại hệ thống AI khác nhau. Một số hệ thống được triển khai hoàn toàn bằng phần mềm, trong khi những hệ thống khác phần lớn được triển khai bằng phần cứng (ví dụ như người máy).

Một định nghĩa có thể chấp nhận được về tính đáng tin cậy là khả năng đáp ứng kỳ vọng của các bên liên quan theo cách thức có thể kiểm chứng được. Định nghĩa này có thể được áp dụng cho một loạt các hệ thống, công nghệ và lĩnh vực ứng dụng AI.

Cũng như đối với tính bảo mật, tính đáng tin cậy đã được hiểu và coi như một yêu cầu phi chức năng xác định các thuộc tính mới nổi của một hệ thống – tức là một tập các đặc điểm cùng với các thuộc tính của chúng – bên trong bối cảnh chất lượng sử dụng. Điều này được chỉ ra trong ISO/IEC 25010 [20].

Ngoài ra, giống như tính bảo mật, tính đáng tin cậy có thể được cải thiện thông qua một quy trình có tổ chức với các kết quả cụ thể có thể đo lường và các chỉ số hiệu năng chính (KPI).

Tóm lại, tính đáng tin cậy được hiểu và xem như một quy trình có tổ chức được tiến hành đi cùng với một yêu cầu (phi chức năng).

Theo UNEP [26], "nguyên tắc phòng ngừa" có nghĩa là ở những nơi có các mối đe dọa nghiêm trọng hoặc tổn hại không thể phục hồi, thì việc cung cấp các sở cứ tin chắc về mặt khoa học về các mối đe dọa và tổn hại đó sẽ không được sử dụng như một lý do để trì hoãn các biện pháp hữu hiệu ngăn ngừa tổn hại. Trong kỹ thuật an toàn, một quy trình để nắm bắt và sau đó xác định kích cỡ, các yêu cầu về "giá trị" của các bên liên quan bao gồm sự hiểu biết về bối cảnh sử dụng của hệ thống, các nguy cơ gây hại và áp dụng "nguyên tắc phòng ngừa" như một kỹ thuật giảm thiểu rủi ro chống lại những hậu quả tiềm ẩn ngoài ý muốn, chẳng hạn như tổn hại đến quyền và tự do của thể nhân, cuộc sống dưới bất kỳ hình thức nào, môi trường, sinh vật hoặc cộng đồng.

Hệ thống AI thường là các hệ thống hiện có được tăng cường khả năng AI. Trong trường hợp này, tất cả các phương pháp tiếp cận và mối quan tâm về tính đáng tin cậy của phiên bản cũ của hệ thống tiếp tục được áp dụng cho hệ thống được cải tiến nâng cao. Chúng bao gồm các phương pháp tiếp cận về chất lượng (cả dữ liệu đo đạc và phương pháp đo lường), tính an toàn, nguy cơ tổn hại và các khuôn

khổ quản lý rủi ro (chẳng hạn như các khuôn khổ hiện có về tính bảo mật và tính riêng tư). Các mục từ 5.2 đến 5.5 trình bày các khuôn khổ khác nhau để bối cảnh hóa tính đáng tin cậy của các hệ thống AI.

5.2 Nhận biết các lớp tin cậy

Một hệ thống AI có thể được khái quá hóa là sự hoạt động trong một hệ sinh thái gồm các lớp chức năng. Sự tin cậy được thiết lập và duy trì ở mỗi lớp để hệ thống AI được tin cậy trong môi trường của nó. Ví dụ: báo cáo của ITU-T về cung cấp sự tin cậy [27] giới thiệu ba lớp tin cậy: tin cậy vật lý, tin cậy không gian mạng và tin cậy xã hội trên cơ sở xem xét đến hạ tầng vật lý thu thập dữ liệu (ví dụ: các bộ cảm biến và thiết bị truyền động), hạ tầng công nghệ thông tin lưu trữ dữ liệu và xử lý (ví dụ như đám mây) và ứng dụng đầu cuối (ví dụ: thuật toán ML, hệ thống chuyên gia và ứng dụng cho người dùng đầu cuối).

Xét về lớp tin cậy vật lý, khái niệm này thường đồng nghĩa với sự kết hợp giữa độ tin cậy và độ an toàn vì các số liệu dựa trên phép đo đạc hoặc kiểm tra vật lý. Ví dụ, kiểm soát kỹ thuật một chiếc ô tô làm cho chiếc xe và các cơ cấu bên trong của nó trở nên đáng tin cậy. Trong bối cảnh này, mức độ tin cậy có thể được xác định thông qua mức độ đáp ứng đối với một danh sách kiểm tra. Ngoài ra, một số quy trình như hiệu chuẩn cảm biến có thể đảm bảo tính chính xác của các phép đo và do đó cũng như dữ liệu được tạo ra.

Ở lớp tin cậy không gian mạng, các mối quan tâm thường là các yêu cầu bảo mật cơ sở hạ tầng công nghệ thông tin, chẳng hạn như kiểm soát truy cập và các biện pháp khác để duy trì tính toàn vẹn của hệ thống AI và giữ an toàn cho dữ liệu của nó.

Sự tin tưởng ở lớp ứng dụng đầu cuối của hệ thống AI yêu cầu phần mềm và các thứ khác đi cùng với nó là đáng tin cậy và an toàn. Trong bối cảnh của các hệ thống quan trọng, việc sản xuất phần mềm được đóng khung bởi một bộ các quy trình để xác minh và thẩm định "sản phẩm" [28]. Điều này cũng đúng với các hệ thống AI và còn hơn thế nữa. Với bản chất ngẫu nhiên của các hệ thống AI dựa trên máy học, tính đáng tin cậy cũng ngụ ý rằng sự công bằng trong hành vi của hệ thống là tương ứng với việc không có sự thiên vị không phù hợp.

Hơn nữa, tin cậy xã hội dựa trên cách sống, niềm tin, tính cách v.v.. của một người. Nếu không có hiểu biết rõ ràng sự vận hành bên trong thì các nguyên lý hoạt động của nó sẽ là không minh bạch đối với bộ phận dân chúng không có chuyên môn về kỹ thuật. Trong trường hợp này, việc thiết lập sự tin cậy có thể không phụ thuộc vào xác minh khách quan về hiệu quả của hệ thống AI mà dựa trên sự giảng giải mang tính chủ quan về hành vi quan sát được từ hệ thống AI.

5.3 Áp dụng các tiêu chuẩn chất lượng phần mềm và dữ liệu

Phần mềm có ảnh hưởng quan trọng đến tính đáng tin cậy của một hệ thống AI điển hình. Do đó, việc xác định và mô tả các thuộc tính chất lượng của phần mềm có thể giúp cải thiện tính đáng tin cậy của toàn bộ hệ thống [29]. Những thuộc tính này có thể góp phần nâng cao tin cậy không gian mạng và tin cậy xã hội. Ví dụ, từ góc độ xã hội, tính đáng tin cậy được mô tả bằng khả năng, tính toàn vẹn và nhân đạo [30]. Dưới đây là các ví dụ về cách các thuật ngữ này được diễn giải trong bối cảnh của các hệ

thống AI.

- Khả năng là năng lực của hệ thống AI thực hiện một tác vụ cụ thể (ví dụ: phát hiện khối u trong chuẩn đoán hình ảnh hoặc xác định một người bằng nhận dạng khuôn mặt qua hệ thống giám sát video). Các thuộc tính liên quan đến khả năng bao gồm độ bền vững, an toàn, tin cậy v.v..
- Tính toàn vẹn là sự tôn trọng của hệ thống AI đối với các nguyên tắc đạo đức hoặc sự đảm bảo rằng thông tin sẽ không bị hệ thống AI thao túng theo cách độc hại. Do đó, các thuộc tính của tính toàn vẹn bao gồm tính đầy đủ, chính xác, chắc chắn, nhất quán v.v..
- Tính nhân đạo là mức độ mà hệ thống AI được cho là làm việc tốt, hay nói cách khác là tôn trọng nguyên tắc "không gây hại".

ISO/IEC SQuaRE đề cập đến chất lượng phần mềm thông qua các mô hình và phép đo (ISO/IEC 2501x về mô hình và ISO/IEC 2502x về đo lường) để từ đó có được một danh mục các đặc điểm của chất lượng phần mềm và dữ liệu.

SQuaRE phân biệt giữa các mô hình như sau:

- Chất lượng phần mềm có 8 đặc điểm;
- Chất lượng sử dụng phần mềm, dữ liệu và dịch vụ công nghệ thông tin có 5 đặc điểm được dùng để phân biệt giữa tin cậy không gian mạng và tin cậy xã hội và chỉ rõ những rủi ro có thể xảy ra cần phải giảm thiểu;
- Chất lượng dữ liệu gồm 15 đặc điểm; và
- Chất lượng dịch vụ công nghệ thông tin gồm 8 đặc điểm.

Ví dụ, theo tiêu chuẩn ISO/IEC 25010 [20], các yêu cầu xã hội mới xuất hiện gần đây thuộc phạm trù "không rủi ro". Theo [20] không rủi ro được hiểu là "mức độ mà một sản phẩm hoặc hệ thống giảm thiểu rủi ro tiềm tàng đến tình trạng kinh tế, tính mạng con người, sức khỏe hoặc môi trường".

ISO/IEC 25010 là một phần của hệ tiêu chuẩn quốc tế SquaRE, đưa ra một mô hình gồm các đặc điểm chính và các đặc điểm phụ về chất lượng phẩm phần mềm và chất lượng sử dụng phần mềm. ISO/IEC 25012 [19] cũng là một phần của SquaRE xác định mô hình chất lượng dữ liệu chung cho dữ liệu xử lý ở định dạng có cấu trúc trong hệ thống máy tính. ISO/IEC 25012 tập trung vào chất lượng của dữ liệu như một phần của hệ thống máy tính và xác định các đặc tính chất lượng dữ liệu đích được sử dụng bởi con người và hệ thống.

SQuaRE được phát triển cho các hệ thống phần mềm lưu trữ dữ liệu truyền thống có cấu trúc và xử lý nó bằng quan hệ logic tường minh. ISO/IEC 25012 mô tả mô hình chất lượng dữ liệu bằng việc sử dụng 15 đặc điểm khác nhau như độ chính xác, tính đầy đủ, khả năng tiếp cận, khả năng truy xuất nguồn gốc, tính di động v.v..

Việc đo lường cả đặc tính chất lượng hệ thống và chất lượng dữ liệu là một thách thức. Mô hình chất lượng dữ liệu trong ISO/IEC 25012 không giải quyết đầy đủ tất cả các đặc điểm về bản chất dữ liệu của hệ thống AI. Ví dụ học sâu là một cách tiếp cận để tạo ra các biểu diễn phân cấp phong phú thông qua

việc đào tạo mạng nơ-ron với nhiều lớp ẩn dựa vào lượng lớn dữ liệu [31]. Ngoài ra, mô hình chất lượng dữ liệu của hệ thống AI còn cần phải xem xét ở các đặc điểm khác hiện được mô tả trong ISO/IEC 25012, chẳng hạn như sự thiên vị trong dữ liệu được sử dụng để phát triển hệ thống AI.

Để bao quát đầy đủ hơn các hệ thống AI và dữ liệu của nó, có thể cần phải mở rộng hoặc sửa đổi các tiêu chuẩn hiện có về các đặc điểm và yêu cầu đối với phát triển phần mềm và hệ thống truyền thống được mô tả trong ISO/IEC 25010, mô hình chất lượng dữ liệu trong ISO/IEC 25012.

5.4 Áp dụng quản lý rủi ro

Quản lý rủi ro là một quá trình phòng ngừa nhằm đảm bảo rằng sản phẩm AI hoặc dịch vụ AI "theo thiết kế" có tính đáng tin cậy trong suốt vòng đời của nó. Quy trình chung quản lý rủi ro được định nghĩa trong ISO 31000: 2018 [14] đề cập đến việc xác định các bên liên quan, tài sản và giá trị dễ bị tổn thương, đánh giá rủi ro liên quan đến hậu quả hoặc tác động của chúng, đưa ra quyết định xử lý rủi ro tối ưu dựa trên các mục tiêu cần đạt và khả năng chấp nhận rủi ro của tổ chức. Rủi ro theo ISO 31000 [14] được định nghĩa là "ảnh hưởng của sự không chắc chắn đối với các mục tiêu", trong đó ảnh hưởng là sự sai lệch so với dự kiến và nó được đo lường hoặc đánh giá theo khả năng xảy ra sự kiện không mong muốn và mức độ tác động có thể xảy ra từ các bên liên quan. Quản lý rủi ro đặc biệt phù hợp với các công nghệ mới mà những cái chưa lường trước nhiều hơn những cái có thể lường trước được. Nó cũng phù hợp để đối phó với các tình huống tiềm ẩn rủi ro, chẳng hạn như đối phó với lỗi của con người và các tấn công độc hại. Hơn nữa, quản lý rủi ro giúp đối phó với sự không chắc chắn trong các lĩnh vực chưa có các phép đo lường chất lượng được công nhận. Những đặc điểm nói trên phổ biến trong các hệ thống AI khiến chúng đặc biệt thích hợp để quản lý rủi ro.

Các khái niệm chủ yếu trong ISO 31000 được trình bày ở đây để chỉ ra cách chúng có thể được áp dụng cho các hệ thống AI. Quản lý rủi ro là một quá trình phòng ngừa nhằm đảm bảo rằng sản phẩm AI hoặc dịch vụ AI ""theo thiết kế" có tính đáng tin cậy trong suốt vòng đời của nó. Quy trình chung quản lý rủi ro được định nghĩa trong ISO 31000: 2018 [14] đề cập đến việc xác định các bên liên quan, tài sản và giá trị dễ bị tổn thương, đánh giá rủi ro liên quan đến hậu quả hoặc tác động của chúng, đưa ra quyết định xử lý rủi ro tối ưu dựa trên các mục tiêu cần đạt và khả năng chấp nhận rủi ro của tổ chức. Rủi ro theo ISO 31000 [14] được định nghĩa là "ảnh hưởng của sự không chắc chắn đối với các mục tiêu", trong đó ảnh hưởng là sự sai lệch so với dự kiến và nó được đo lường hoặc đánh giá theo khả năng xảy ra sự kiện không mong muốn và mức độ tác động có thể xảy ra từ các bên liên quan. Quản lý rủi ro đặc biệt phù hợp với các công nghệ mới mà những cái chưa lường trước nhiều hơn những cái có thể lường trước được. Nó cũng phù hợp để đối phó với các tình huống tiềm ẩn rủi ro, chẳng hạn như đối phó với lỗi của con người và các tấn công độc hại. Hơn nữa, quản lý rủi ro giúp đối phó với sự không chắc chắn trong các lĩnh vực chưa có các phép đo lường chất lượng được công nhận. Những đặc điểm nói trên phổ biến trong các hệ thống AI khiến chúng đặc biệt thích hợp để quản lý rủi ro. (hoặc nhận dạng các nguồn gốc rủi ro) một cách hữu hình hơn. Các mục tiêu kiểm soát thường là tính dễ tổn thương, cạm

bẫy hoặc các mối đe dọa đã được dự đoán[1]. Đối với các hệ thống AI, chúng sẽ bao gồm (nhưng không giới hạn) các thách thức về trách nhiệm giải trình, các mối đe dọa bảo mật và quyền riêng tư mới, đặc điểm kỹ thuật không phù hợp, triển khai có thiếu sót, sử dụng không đúng và các nguồn phát sinh sự thiên vị khác nhau. Đối với mỗi mục tiêu kiểm soát đã xác định, có thể xác định được một tập các biện pháp kiểm soát (hoặc giảm thiểu). Đối với các hệ thống AI, chúng sẽ bao gồm (nhưng không giới hạn) ở:

- Các phương pháp tiếp cận tính minh bạch;
- Các biện pháp kiểm soát an ninh mới;
- Chính sách quản lý mới;
- Các mối quan tâm đến độ bền vững và khả năng phục hồi;
- Các quan tâm liên quan đến lựa chọn và cấu hình các thuật toán ML;
- Các quan tâm đến dữ liệu; và
- Các quan tâm liên quan đến khả năng kiểm soát của hệ thống.

Quá trình quản lý rủi ro thực hiện từng biện pháp kiểm soát và chỉ ra một tập hợp các chỉ dẫn (hoặc biện pháp) phù hợp với chính sách của tổ chức và các tình huống. Một khi khuôn khổ quá trình quản lý rủi ro được tạo lập thì việc thực hiện đúng và triển khai chính xác phải được kiểm tra, xem xét và cải tiến liên tục bằng các phương pháp đánh giá và đo lường khác nhau bao gồm (nhưng không giới hạn ở) các số liệu đo được về hiệu năng thuật toán và kết quả thử nghiệm tại hiện trường.

5.5 Các phương pháp tiếp cận có sự trợ giúp của phần cứng

Các hệ thống học máy điển hình (gồm cả huấn luyện và sử dụng) được triển khai trên các nền tảng phổ dụng có sẵn là không đáng tin cậy và có thể ảnh hưởng đến sự hoạt động chính xác của hệ thống. Ví dụ các ứng dụng học máy thường triển khai trên môi trường đám mây nhiều người thuê. Các cơ chế phần cứng hỗ trợ giảm thiểu tấn công bề mặt bằng cách cung cấp các môi trường thực thi tin cậy (TEE) để bảo vệ tính bí mật và tính toàn vẹn của cả dữ liệu và các bước tính toán, cũng như cho cả hoạt động tập huấn và sử dụng.

TEE được sử dụng để bảo vệ mã nguồn và dữ liệu được chọn khỏi bị tiết lộ hoặc sửa đổi bằng việc cung cấp sự cách ly cưỡng bức bằng phần cứng đối với các chương trình hoặc các khu vực hoạt động cần được bảo vệ nhằm tăng cường bảo mật ngay cả trên các nền tảng bị xâm hại. Sử dụng môi trường thực thi đáng tin cậy cho phép các nhà phát triển có thể bảo vệ mô hình học máy trong suốt vòng đời của nó (ví dụ như huấn luyện và sử dụng nó), cách thức hiệu quả là coi mô hình đó như dữ liệu hoặc tài sản trí

---

[1] Lưu ý rằng việc ánh xạ giữa các mục tiêu của tổ chức và các mục tiêu kiểm soát có thể không là một – một. Đối với các hệ thống phức tạp (chẳng hạn như hệ thống AI), việc đạt được từng mục tiêu của tổ chức thường đòi hỏi phải đạt được nhiều mục tiêu kiểm soát nhất định.

tuệ được bảo vệ khi cần thiết. TEE thực thi tính bảo mật và tính toàn vẹn của bộ nhớ được sử dụng bởi khối lượng công việc của ML (thường sử dụng là công cụ mã hóa bộ nhớ) ngay cả khi hiện diện sự can thiệp của các phần mềm độc hại ở các phân lớp phần mềm hệ thống.

6 Các bên liên quan

6.1 Các khái niệm chung

Tài liệu này chấp nhận một định nghĩa mở rộng về các bên liên quan từ ISO/IEC 38500 [17] để bổ sung cho việc thừa nhận vai trò của các cá nhân và tổ chức:

- Thừa nhận một nhóm người là một loại hình bên liên quan, điều quan trọng là hiểu được các quan điểm tập thể được chia sẻ bởi một nhóm các cá nhân không cấu thành một tổ chức, nghĩa là không có một sự quản lý chung để đại diện cho nhóm đó; và
- Sự bao hàm loại hình các bên liên quan có thể bị tác động bởi hệ thống, sự bao hàm đó nhiều khi mở rộng ra ngoài phạm vi những bên cần thiết hoặc được hệ thống mong đợi, điều này ngụ ý rằng hệ thống cần phải có những hiểu biết trước về những điều đó.

Định nghĩa mở rộng này rất quan trọng để xem xét, xác định các bên liên quan vì có thể trong số đó có những đối tượng không biết về sự tồn tại, mục tiêu hoặc năng lực của hệ thống.

Định nghĩa của thuật ngữ "tài sản" trong ISO/IEC 27000: 2018 [1] là không đủ khi xem xét mối quan hệ với các bên liên quan như đã thảo luận ở trên. Thay vào đó, bằng cách sử dụng thuật ngữ tài sản, tiêu chuẩn này đề cập đến "bất kỳ thứ gì có giá trị đối với các bên liên quan". Điều này mở rộng giả định của tài liệu tham khảo [1] cho rằng chỉ các tổ chức mới sở hữu tài sản có giá trị mà trong nhiều trường hợp đó lại là các cá nhân hoặc nhóm người.

Các giá trị, trong ngữ cảnh của tiêu chuẩn này không giới hạn ở các tổ chức (theo tài liệu tham khảo [23]), nhưng bao gồm niềm tin mà bất kỳ bên liên quan nào "dựa vào và các chuẩn mực tìm kiếm để tuân thủ".

Với khả năng vận hành các hệ thống AI theo kiểu linh hoạt và năng động, các phương pháp tiếp cận tính đáng tin cậy của AI cần tập trung vào cả việc tăng cường và duy trì sự tin cậy. Điều này có thể đạt được bằng sử dụng các định nghĩa để cung cấp bối cảnh rõ ràng với các đặc điểm cụ thể của một AI đáng tin cậy, chẳng hạn như một sự thay đổi về bối cảnh có thể kích hoạt một hoạt động đánh giá lại quan trọng đối với một đặc tính đã được trạng thái hóa [32]. Theo nghĩa này, sẽ là không đủ nếu chỉ đơn giản đề cập đến "AI đáng tin cậy" mà phải chỉ rõ ai tin cậy ai trong những khía cạnh nào của phát triển và sử dụng AI. Do đó, việc bối cảnh hóa các bên liên quan như vậy có thể áp dụng để xem xét các đặc điểm AI đáng tin cậy như là tính minh bạch (xem 9.2), tính diễn giải (xem 9.3) và khả năng kiểm soát (xem 9.4). Việc ngữ cảnh hóa yêu cầu phải xác định rõ các bên liên quan và hiểu rõ ràng về sự tham gia của họ tại các điểm khác nhau trong vòng đời và chuỗi giá trị của hệ thống AI.

Các bên liên quan khác nhau có thể có quan điểm khác nhau về tầm quan trọng đối với các đặc điểm được đề xuất khác nhau về một AI đáng tin cậy. Do đó việc tiêu chuẩn hóa các thuật ngữ và khái niệm

khung cho AI đáng tin cậy cho phép giao tiếp một cách rõ ràng giữa các bên liên quan khác nhau, để từ đó có thể hiểu được những khác biệt về quan điểm và giải quyết những khác biệt này. Giao tiếp với các bên liên quan như vậy sẽ giải quyết những vấn đề như:

- Các bên liên quan khác nhau có thể bị ảnh hưởng như thế nào bởi công nghệ AI được triển khai trong một sản phẩm hoặc dịch vụ;
- Bất kỳ một tài sản nào được định giá bởi các bên liên quan khác nhau được sử dụng hoặc bị ảnh hưởng như thế nào bởi việc sử dụng AI trong một sản phẩm hoặc dịch vụ;
- Việc sử dụng AI trong một sản phẩm hoặc dịch vụ có liên quan như thế nào đến các giá trị do các bên liên quan khác nhau nắm giữ.

6.2 Loại hình

Hiện vẫn chưa chưa có sự đồng thuận rõ ràng về loại hình các bên liên quan đối với việc sử dụng AI trong các sản phẩm hoặc dịch vụ. Trong kinh doanh, lý thuyết về các bên liên quan [33] nhấn mạnh lợi ích của cách tiếp cận ra quyết định vượt ra ngoài nghĩa vụ được ủy thác của ban lãnh đạo để tạo ra lợi nhuận cho các cổ đông và xem xét lợi ích của các bên liên quan khác trong một tổ chức bao gồm: nhân viên, khách hàng, ban giám đốc, nhà cung cấp, chủ nợ, chính phủ và các cơ quan quản lý, xã hội nói chung và môi trường tự nhiên (như một đại diện cho các thế hệ tương lai).

Trong bối cảnh của AI, chúng ta có thể xem xét các loại hình bên liên quan như vậy đối với các vai trò riêng biệt sau đây trong chuỗi giá trị AI (lưu ý rằng một bên liên quan duy nhất có thể đảm nhận một vài vai trò như vậy):

- Nguồn dữ liệu: tổ chức hoặc cá nhân cung cấp dữ liệu sử dụng để huấn luyện một hệ thống AI;
- Nhà phát triển hệ thống AI: tổ chức hoặc cá nhân thiết kế, phát triển và huấn luyện một hệ thống AI;
- Nhà sản xuất AI: tổ chức hoặc cá nhân thiết kế, phát triển, thử nghiệm và triển khai sản phẩm hoặc dịch vụ sử dụng ít nhất một hệ thống AI;
- Người sử dụng AI: tổ chức hoặc cá nhân tiêu thụ sản phẩm hoặc dịch vụ sử dụng ít nhất một hệ thống AI;
- Nhà phát triển công cụ và phần sụn AI: một tổ chức hoặc cá nhân thiết kế và phát triển các công cụ AI và các khối tạo dựng AI được huấn luyện trước;
- Cơ quan kiểm tra và đánh giá: một tổ chức hoặc một cá nhân cung cấp thử nghiệm độc lập và có thể thực hiện chứng nhận;
- Cộng đồng mở rộng trong đó hệ thống AI được triển khai (vì ngay cả một hệ thống AI chính xác cũng có thể dẫn đến việc xác nhận các bất bình đẳng hiện hữu);
- Các hiệp hội đại diện cho quan điểm của các nhóm cá nhân;
- Các tổ chức quản lý, giám sát và nghiên cứu việc sử dụng AI bao gồm các chính phủ của các quốc gia và các tổ chức quốc tế, chẳng hạn như Quỹ Tiền tệ quốc tế (IMF).

6.3 Tài sản

Có thể đặc điểm hóa các bên liên quan bằng tài sản mà họ có vai trò đánh giá hoặc bị ảnh hưởng bởi sử dụng AI trong sản phẩm hoặc dịch vụ.

Các tài sản hữu hình đối với AI có thể bao gồm:

- Dữ liệu sử dụng để huấn luyện hệ thống AI;
- Một hệ thống AI được huấn luyện;
- Một sản phẩm hoặc dịch vụ sử dụng một hoặc nhiều hệ thống AI;
- Dữ liệu được sử dụng để kiểm tra hành vi liên quan đến AI của một sản phẩm hoặc dịch vụ;
- Dữ liệu cung cấp cho hoạt động của sản phẩm hoặc dịch vụ mà theo đó các quyết định dựa trên AI được đưa ra;
- Tài nguyên máy tính và phần mềm sử dụng để huấn luyện, thử nghiệm và vận hành các hệ thống AI;
- Nguồn nhân lực với các kỹ năng để:
  - Huấn luyện, thử nghiệm và vận hành các hệ thống AI;
  - Phát triển phần mềm sử dụng trong hoặc cho các tác vụ đó; và / hoặc
  - Tạo, chú giải hoặc chọn dữ liệu cần thiết cho huấn luyện AI.

Tài sản ít hữu hình hơn bao gồm:

- Danh tiếng và sự tin cậy được tạo lập, bên liên quan tới phát triển, thử nghiệm hoặc vận hành hệ thống AI; hoặc dịch vụ, sản phẩm sử dụng;
- Thời gian, ví dụ như thời gian mà người dùng sản phẩm hoặc dịch vụ sử dụng AI tiết kiệm được hoặc thời gian lãng phí để xử lý đề xuất không phù hợp từ hệ thống AI;
- Các kỹ năng, có thể trở nên ít được coi trọng hơn do chức năng tự động hóa có trong hệ thống AI;
- Quyền tự trị có thể được tăng cường hoặc bị xói mòn phụ thuộc vào việc hệ thống AI cung cấp thông tin liên quan đến tác vụ có hiệu quả hay không. Ví dụ như quảng cáo hoặc thông điệp chính trị có nội dung thuyết phục được truyền đạt có chủ ý đến các cá nhân hoặc một nhóm cá nhân bằng việc sử dụng hồ sơ trong AI.

6.4 Giá trị

Các bên liên quan có thể có quan điểm khác nhau về các đặc điểm thích hợp đối với một hệ thống AI đáng tin cậy dựa trên các giá trị khác nhau mà họ tuân thủ hoặc tìm cách quan sát. Một số đề xuất về AI đáng tin cậy dựa trên một bộ giá trị cụ thể được tạo lập trong một chính sách cụ thể, chẳng hạn như tài liệu làm việc của nhóm chuyên gia cấp cao của Ủy ban Châu Âu về AI đáng tin cậy [34] đề xuất các nguyên tắc dựa trên Hiến chương các quyền cơ bản của Châu Âu. Các quan điểm khác nhau về AI đáng tin cậy cũng có thể rút ra từ thế giới quan hoặc hệ thống giá trị đạo đức khác nhau. Mức độ liên quan và tác động của các thế giới quan khác nhau, chẳng hạn như đạo đức phương Tây, phật giáo, triết lý

Ubuntu, thần đạo trong AI vẫn gần như chưa được khám phá [35]. Nói chung việc thiết kế mang tính nhạy cảm với các giá trị [36] thì sự hiểu biết rõ ràng những khác biệt về giá trị này là điều cần thiết trong việc truyền đạt các đặc điểm AI đáng tin cậy ở cấp độ toàn cầu.

7 Nhận biết các mối quan tâm cấp cao

7.1 Trách nhiệm, trách nhiệm giải trình và quản trị

Phát triển và ứng dụng các hệ thống AI là việc ứng dụng công nghệ thông tin trong môi trường nhiều bên liên quan. Để xây dựng và duy trì niềm tin trong một môi trường như vậy, điều quan trọng là phải xác định trách nhiệm và trách nhiệm giải trình giữa các bên liên quan. Các hệ thống AI có thể tồn tại trong cả một chuỗi giá trị thương mại quốc tế phức tạp và trên các khuôn khổ xã hội xuyên quốc gia. Điều quan trọng là tất cả các bên liên quan chia sẻ sự hiểu biết về trách nhiệm mà họ thực hiện đối với các bên liên quan khác và cách họ chịu trách nhiệm đối với những bổn phận đó. Một trong những lý do chính để có sự thống nhất về một khuôn khổ như vậy là khả năng xác định các điểm ra quyết định trong suốt vòng đời của hệ thống AI.

Trong một tổ chức, trách nhiệm đối với các quyết định và trách nhiệm giải trình về kết quả của các quyết định đó thường được quy định trong một khuôn khổ quản trị. ISO/IEC 38500 [17] hướng dẫn những người ra quyết định cấp cao trong tổ chức hiểu và tuân thủ các quy định pháp luật, quy định quản lý và nghĩa vụ đạo đức trong sử dụng công nghệ thông tin. Nó xác định các tác vụ đánh giá, chỉ đạo và giám sát các khía cạnh về công nghệ thông tin trong việc thực thi các nguyên tắc về trách nhiệm, chiến lược, thu nhận, hiệu năng, sự tuân thủ và hành vi của con người. ISO/IEC 38505 [37] áp dụng mô hình quản trị công nghệ thông tin này vào quản trị dữ liệu. Nó không đề cập đến các lỗ hổng của AI nhưng đề cập đến một số vấn đề liên quan liên quan đến định hướng dữ liệu AI, chẳng hạn như học máy. Có thể tồn tại cơ hội gắn kết hơn nữa mô hình quản trị công nghệ thông tin theo ISO/IEC 38500 với những yêu cầu cần thiết đối với hệ thống AI đáng tin cậy, đặc biệt là những vấn đề liên quan đến sự tương tác của chúng với các khuôn khổ ra quyết định mang tính xã hội khác liên quan đến sử dụng chúng trong các hệ thống tự trị. Thuật ngữ "hệ thống tự trị" được sử dụng trong tiêu chuẩn này để chỉ bất kỳ loại hệ thống nào hoạt động tương đối độc lập và không giới hạn ở khái niệm ô tô hoặc "máy móc".

Ví dụ: bác sĩ có thể sử dụng AI để cải thiện các chẩn đoán của mình. Bác sĩ chịu trách nhiệm về chẩn đoán của mình vì là một chuyên gia có trình độ trong lĩnh vực này, do đó là người chịu trách nhiệm phân tích kết quả đầu đối với một hệ thống AI ra quyết định chuẩn đoán. Một ví dụ khác, một người dùng đầu cuối nộp đơn xin việc và người ấy không thông hiểu về hệ thống AI và sẽ khó để hiểu lý do tại sao đơn xin việc của mình bị từ chối. Trong trường hợp như vậy chuỗi trách nhiệm cần được xác định rõ ràng. Để đạt được tính tin cậy của các hệ thống tự trị do AI điều khiển thì cần phải giải quyết các trách nhiệm và trách nhiệm giải trình trong trường hợp hệ thống tự quản trị bị lỗi [38] [40]. Điều này cho phép các bên liên quan có sự liên đới chịu trách nhiệm pháp lý nếu một hệ thống tự quản gây ra thiệt hại. Nhóm công tác của châu Âu về đạo đức trong khoa học và công nghệ mới [41] đã nhấn mạnh trong tuyên bố năm 2018 của họ, rằng một hệ thống do AI điều khiển không thể tự trị theo nghĩa pháp lý và cần phải thiết lập

một khuôn khổ trách nhiệm và trách nhiệm pháp lý rõ ràng để có thể truy cứu bất kỳ thiệt hại nào gây ra bởi hoạt động của bất kỳ hệ thống tự trị nào.

7.2 Tính an toàn

Tính an toàn là một khía cạnh quan trọng của tính đáng tin cậy. Do đó việc xem xét các khía cạnh về tính an toàn được ưu tiên cao. Thông thường rủi ro gây hại của hệ thống được nhận thức càng cao thì yêu cầu về tính đáng tin cậy càng cao.

Hệ thống AI cũng như bất kỳ hệ thống nào khác, được kỳ vọng sẽ không gây ra bất kỳ tác hại không có chủ ý nào. Điều này không chỉ bao gồm tác hại hữu hình (ví dụ, đối với sức khỏe của sinh vật, tài sản và môi trường vật chất) mà còn cả tác hại vô hình (ví dụ đối với môi trường xã hội và văn hóa).

ISO/IEC Hướng dẫn 51 [10] nêu rõ: "Tất cả các sản phẩm và hệ thống đều có các nguy cơ và do đó luôn tồn tại rủi ro ở các mức độ nào đó. Tuy nhiên, rủi ro liên quan đến những nguy cơ đó cần được giảm xuống mức có thể chấp nhận được. Tính an toàn đạt được bằng cách giảm rủi ro đến mức có thể chấp nhận, được định nghĩa trong Hướng dẫn này là rủi ro có thể chấp nhận được". Các hệ thống AI cung cấp một mức độ tự chủ nhất định thường được xem là coi trọng tính an toàn hơn. Tuy nhiên các nguy cơ và rủi ro phụ thuộc vào ứng dụng và không nhất thiết liên quan trực tiếp đến mức độ tự chủ (ví dụ phương tiện tự hành đường bộ so với máy hút bụi tự hành dùng cho gia đình).

Vòng đời hoàn chỉnh của một hệ thống AI từ thiết kế đến xử lý đều phải được xem xét các khía cạnh về tính an toàn. Đối với ứng dụng hệ thống AI, các mục đích sử dụng và lạm dụng việc sử dụng cần phải được nhận lường trước một cách hợp lý; phải xem xét cẩn thận môi trường mà nó được sử dụng cũng như các công nghệ được sử dụng. ISO/IEC Hướng dẫn 51 [10] định nghĩa "lạm dụng có thể lường trước được một cách hợp lý" là việc sử dụng sản phẩm hoặc hệ thống theo cách không do nhà cung cấp đưa ra, nhưng đó là kết quả của hành vi con người mà hệ thống đã lường trước. Hệ thống AI có khả năng đưa ra rủi ro xuất phát từ các lỗ hổng cụ thể của AI. Điều này sẽ dẫn đến các biện pháp mới để giảm rủi ro xuống mức có thể chấp nhận căn cứ vào các hành vi cụ thể của AI, chẳng hạn như hoạt động ra quyết định không có tính minh bạch hoặc không được xác định trước. Đối với một hệ thống cụ thể, không chỉ xét riêng về thành phần AI mà tất cả các công nghệ được sử dụng cũng như sự tương tác giữa chúng đều phải được xem xét cẩn thận. ISO/IEC Hướng dẫn 51 [10] đưa ra một chỉ dẫn chung để xác định các rủi ro có thể chấp nhận được.

8 Tính dễ bị tổn thương, các mối đe dọa và thách thức

8.1 Yêu cầu chung

Mục này mô tả tính dễ bị tổn thương tiềm ẩn của các hệ thống AI và các mối đe dọa liên quan đến chúng.

Tính dễ bị tổn thương được ISO/IEC 27000 [1] định nghĩa là điểm yếu của một tài sản hoặc điều khiển có thể có thể bị lợi dụng bởi một hoặc nhiều mối đe dọa. Các mối đe dọa được ISO/IEC 27000 [1] định nghĩa là nguyên nhân tiềm ẩn của sự cố không mong muốn, có thể dẫn đến tổn hại cho hệ thống hoặc tổ chức.

Các bên liên quan khác nhau sử dụng các thuật ngữ khác nhau để mô tả khái niệm "tính dễ bị tổn thương". Chúng bao gồm các nguồn rủi ro, cạm bẫy, nguồn gốc các hư hỏng, nguyên nhân có tính căn nguyên, các thách thức.

Tính dễ bị tổn thương liên quan đến việc sử dụng học máy trong các hệ thống AI. Chúng bao gồm sự phụ thuộc vào dữ liệu, tính không rõ ràng của các mô hình ML và tính không thể đoán trước. Việc sử dụng dữ liệu có thể dẫn đến các mối đe dọa mới về bảo mật và sự thiên vị.

Những thách thức liên quan đến việc thiếu "các phương pháp tốt nhất" cho việc thiết kế, phát triển và triển khai các hệ thống AI dẫn đến việc làm gia tăng hoặc trầm trọng thêm tính dễ bị tổn thương và mối đe dọa hiện hữu.

Một số mối đe dọa nhất định nào đó phát sinh do sự hiểu biết không đầy đủ khả năng công nghệ của các hệ thống AI và việc sử dụng chúng không hợp lý bởi các bên liên quan khác nhau.

Các mục từ 8.2 đến 8.10 mô tả chi tiết hơn về tính dễ bị tổn thương, các mối đe dọa và thách thức tiềm ẩn khác của hệ thống AI.

8.2 Các mối đe dọa bảo mật cụ thể của AI

8.2.1 Yêu cầu chung

Sự phát triển của AI mang lại cả ưu và nhược điểm đối với bảo mật số. Một mặt các công nghệ AI có thể được sử dụng để lập hồ sơ những kẻ tấn công và các hoạt động độc hại của chúng để từ đó đưa ra các giải pháp bảo mật để chống lại chúng. Mặt khác công nghệ tiên tiến được phát triển bằng cách sử dụng AI và học máy có thể bị lạm dụng cho các mục đích xấu. Ví dụ AI có thể được sử dụng để dò đoán mật khẩu và xâm phạm tài khoản trong môi trường số.

Ngoài các mối đe dọa bảo mật công nghệ thông tin phổ biến đối với hầu hết các hệ thống (ví dụ lỗi phần mềm, cửa sau của phần cứng, vi phạm bảo mật dữ liệu), một số hệ thống AI nhất định, chẳng hạn như hệ thống học máy có thể dễ bị tấn công bởi các mối đe dọa bảo mật đặc biệt hoặc có mục đích. Những mối đe dọa như vậy bao gồm [43]:

- Đầu độc dữ liệu dẫn đến hệ thống AI bị trục trặc;
- Tấn công đối nghịch hòng lạm dụng các hệ thống AI lành tính; và
- Đánh cắp mô hình.

8.2.2 Đầu độc dữ liệu

Trong các cuộc tấn công đầu độc dữ liệu, những kẻ tấn công cố tình tác động đến dữ liệu huấn luyện để thao túng kết quả của một mô hình dự đoán. Các cuộc tấn công đầu độc dữ liệu có mục đích làm cho máy chủ huấn luyện ra các mô hình tồi mà không biết là đang bị những cuộc tấn công đó. Kiểu tấn công này đặc biệt thích hợp trong trường hợp mô hình được tiếp xúc với Internet ở chế độ trực tuyến, tức là mô hình liên tục được cập nhật bằng cách học hỏi từ dữ liệu mới. Việc lọc dữ liệu huấn luyện một cách thích hợp có thể giúp phát hiện và lọc ra dữ liệu bất thường và do đó giảm thiểu các thiệt hại có thể xảy

ra.

8.2.3 Tấn công đối nghịch

Một mối đe dọa bảo mật cụ thể liên quan đến các hệ thống AI là cuộc tấn công đối nghịch vào các hệ thống học máy. Tấn công đối nghịch bao gồm việc cung cấp dữ liệu đầu vào có sự xáo trộn nhẹ đối với một mô hình. Ví dụ như sửa đổi dữ liệu của một biển báo giao thông nhằm đánh lừa hệ thống phân loại của xe tự hành để nhận dạng sai đối với dữ liệu đầu vào này.

Những tiến bộ ấn tượng gần đây trong lĩnh vực ML, đặc biệt là trong lĩnh vực học sâu đã khiến các công ty ngày càng quan tâm đến việc áp dụng các thuật toán ML trong lĩnh vực an toàn và bảo mật. Ví dụ như việc tích hợp phân đoạn ngữ nghĩa dựa trên mạng nơ-ron phức hợp (CNN) vào ô tô tự hành hoặc mạng nơ-ron vào chẩn đoán y tế. Rõ ràng những bối cảnh ứng dụng mới này có yêu cầu cực kỳ cao về đảm bảo chất lượng. Nhưng cho đến nay những giải pháp trên vẫn chưa chứng minh đạt được độ chính xác cao cho dù các tập dữ liệu huấn luyện, kiểm tra và thẩm định được chuẩn bị một cách kỹ lưỡng.

Bên cạnh việc xử lý các trường hợp phổ biến về phân phối dữ liệu, điều cần thiết là mô-đun ML được huấn luyện phải có khả năng xử lý các trường hợp bất thường, thậm chí với cả trường hợp các điểm đầu vào bị nhiễm độc [44].

Các công bố của Szegedy và cộng sự [45], cũng như của Goodfellow và công sự [46] chỉ ra rằng các thuật toán ML về thị giác máy tính, đặc biệt đối với các mạng nơ-ron sâu có khả năng dễ dàng bị tấn công bởi các mẫu dữ liệu lừa đảo (hoặc làm nhiễu loạn).

Những hành vi tấn công đối nghịch này tạo ra sự nhiễu loạn và thường được xem là hành vi lỗi của mô-đun ML khi bổ sung dữ liệu đầu theo quy trình. Hơn nữa, những nhiễu loạn này thường khó phát hiện hoặc thậm chí không thể nhận thấy bằng mắt thường. Tính chất không thể nhận thấy này không chỉ là thách thức việc mong muốn triển khai các hệ thống AI trong các ngành, lĩnh vực coi trọng về an toàn và bảo mật, mà nó còn đề cập đến khác biệt quan trọng giữa quá trình xử lý thông tin theo giác quan con người và trong mạng nơ-ron nhân tạo [47]. Đã có rất nhiều phương án (chiến lược phòng thủ) khác nhau đã được công bố [48][49] để khắc phục, đối phó với tính dễ bị tổn thương nói trên. Nó trở thành một cuộc chạy đua giữa bên tấn công và bên phòng thủ. Các biện pháp phòng thủ mới được công bố trước một loạt các cuộc tấn công hiện có thường trở nên vô dụng chỉ sau vài tuần do bên tấn công tạo ra các cuộc tấn công mạnh hơn [50]. Các kiểu tấn công như vậy khó có thể đúc kết thành bài học, kinh nghiệm được vì cần phải hiểu biết rõ về cấu trúc liên kết mạng nội bộ bên trong, mà những thành phần này thường bị che khuất khi hệ thống đưa vào hoạt động. Nhưng dù sao vẫn phải đảm bảo rằng chúng được xem xét một cách nghiêm túc từ góc độ tính đáng tin cậy.

8.2.4 Đánh cắp mô hình

Nó ảnh hưởng đến cả tính bảo mật và quyền riêng tư, các cuộc tấn công đánh cắp mô hình sử dụng để "đánh cắp" các mô hình bằng cách sao chép lại các chức năng bên trong của chúng. Điều này được thực hiện bằng cách gửi đến mô hình mục tiêu một số lượng lớn các truy vấn dự báo và sử dụng phản hồi nhận được (các dự đoán) để huấn luyện một mô hình khác.

8.2.5 Các mối đe dọa nhằm vào phần cứng đối với tính bảo mật và tính toàn vẹn

Các ứng dụng học máy thường bị tấn công tương tự như các ứng dụng nhạy cảm khác. Các cuộc tấn công phần mềm và phần cứng điển hình vào các ứng dụng học máy là các cuộc tấn công kỹ thuật số ảnh hưởng đến tính bảo mật, tính toàn vẹn và của dữ liệu và các quá trình tính toán. Các hình thức tấn công khác có thể dẫn đến từ chối dịch vụ (mất tính khả dụng), gây rò rỉ thông tin hoặc dẫn đến hoạt động tính toán không hợp lệ.

Đảm bảo tính bảo mật, tính toàn vẹn của dữ liệu và mã nguồn thông qua cơ chế truyền thống như đảm bảo tính toàn vẹn của bộ nhớ, tính tin cậy của các mô-đun nền tảng (có trong TEE) là cần thiết. Nhưng điều đó không đủ để đảm bảo tính bảo mật và toàn vẹn của mã nguồn và dữ liệu của các công cụ học máy. Do đó thực thi cơ chế ở trên và bắt buộc các chương trình ML tuân thủ logic được lập trình theo dự kiến là quan trọng như nhau. Ví dụ, một cuộc tấn công luồng điều khiển trên một ứng dụng ML có thể đánh bại/phá vỡ hoạt động suy luận của mô hình ML hoặc có thể gây ra huấn luyện không hợp lệ. Loại hình thứ hai rất quan trọng để bắt buộc thực thi tính toàn vẹn về thời gian hoạt động là các cơ chế ngăn chặn các lỗi an toàn bộ nhớ. Các khiếm khuyết logic trong các chương trình có thể bị tận dụng để gây ra lỗi tràn bộ đệm, sử dụng lỗ hổng trong chỉ định bộ nhớ, khai thác vượt quá giới hạn có thể dẫn đến lỗi hoạt động của các ứng dụng ML.

Cần phải xem xét các mối đe dọa đối với các mô hình thiết bị phức tạp, bao gồm cả bộ tăng tốc phần cứng được sử dụng bởi các ứng dụng học máy. Trong rất nhiều trường hợp các bộ tăng tốc hoặc thiết bị có thể bị mô phỏng hoặc giả lập, các ứng dụng trên đám mây có thể lợi dụng từ việc các thiết bị trực tiếp dùng các ứng dụng này. Tuy nhiên, việc xác minh (thông qua chứng thực) các thiết bị giúp đảm bảo rằng thiết bị có khả năng duy trì các yêu cầu về quyền riêng tư và bảo mật của các ứng dụng ML. Khả năng quản lý bộ nhớ vào/ra (IO) phần cứng có thể được sử dụng để kết hợp một cách an toàn các thiết bị với khối lượng công việc, bao gồm cả DMA vào bộ nhớ được bảo vệ. Các véc-tơ tấn công trong tương lai cần chú ý bao gồm giả mạo thiết bị, tấn công ánh xạ lại thời gian chạy và tấn công man-in-the-middle.

8.3 Các mối đe dọa quyền riêng tư điển hình trong AI

8.3.1 Yêu cầu chung

Sự phát triển các thuật toán AI và sử dụng dữ liệu lớn đã cung cấp các giải pháp tinh tế trong nhiều lĩnh vực. Nhiều kỹ thuật AI (ví dụ học sâu) phụ thuộc vào dữ liệu lớn vì độ chính xác của chúng một phần trông cậy vào lượng dữ liệu mà chúng sử dụng. Việc lạm dụng hoặc tiết lộ dữ liệu, đặc biệt là dữ liệu cá nhân và dữ liệu nhạy cảm (ví dụ hồ sơ sức khỏe) có thể gây tác động có hại cho chủ thể dữ liệu. Do đó bảo vệ quyền riêng tư đã trở thành mối quan tâm chính yếu trong AI và phân tích dữ liệu lớn. Thách thức bắt đầu từ giai đoạn đầu của vòng đời dữ liệu (tức là thu thập và chia sẻ dữ liệu giữa các thực thể khác nhau) đến giai đoạn cuối cùng là phân tích dữ liệu và áp dụng các thuật toán AI (ví dụ rủi ro tái nhận dạng sau khi phân tích dữ liệu từ nhiều nguồn).

Những mối đe dọa về quyền riêng tư này có thể dẫn đến những ảnh hưởng tiêu cực đến quyền tự quyết, phẩm giá, quyền tự do và các quyền cơ bản của cá nhân.

8.3.2 Thu thập dữ liệu

Trong quá trình thu thập dữ liệu, mối đe dọa về quyền riêng tư chủ yếu là về lượng dữ liệu cần thu thập cho mục đích nhất định. Một trong những nguyên tắc về quyền riêng tư được giới thiệu trong ISO/IEC 29100 [51] là nguyên tắc tối thiểu hóa dữ liệu. Do sự phụ thuộc của các mô hình học máy vào sự sẵn có của một lượng lớn dữ liệu chất lượng, phong phú và từ nhiều nguồn dữ liệu khác nhau thì việc hạn chế thu thập dữ liệu là một thách thức.

Ngoài ra, mối đe dọa về quyền riêng tư khác đến từ nguy cơ hỏng bộ lưu trữ. Nếu kẻ tấn công xâm phạm bộ nhớ, quyền riêng tư của các chủ thể dữ liệu có thể bị xâm phạm.

8.3.3 Tiền xử lý và mô hình hóa dữ liệu

Tồn tại các mối đe dọa về quyền riêng tư tiềm ẩn trong khi xử lý dữ liệu như sau:

- Suy luận dữ liệu nhạy cảm từ dữ liệu không nhạy cảm bằng cách sử dụng kỹ thuật học máy và AI;
- Dữ liệu cá nhân khả dụng từ nhiều nguồn. Mặc dù dữ liệu được khử nhận dạng, nhưng AI lại có thể nhận dạng dữ liệu bằng cách sử dụng các suy luận dựa trên dữ liệu từ các nguồn khác.

8.3.4 Truy vấn mô hình

Đánh cắp mô hình bằng cách truy vấn mô hình vì những lý do không hợp lệ được mô tả trong 8.2.4. Các cuộc tấn công bảo mật như vậy được thiết kế để làm lộ thông tin bí mật hiện diện trong mô hình ML. Loại tấn công này có thể được sử dụng để tiết lộ thông tin nhạy cảm về các cá nhân, biến nó thành một cuộc tấn công quyền riêng tư.

Các cuộc tấn công như vậy có thể xảy ra trong toàn bộ vòng đời của mô hình, bao gồm cả các giai đoạn phát triển, triển khai và vận hành. Các cuộc tấn công có thể được thực hiện bởi cả tác nhân được phép truy vấn mô hình và những tác nhân khác trước đó đã vi phạm quy định bảo mật để truy cập vào mô hình.

Một mối đe dọa khác liên quan đến việc sử dụng không phù hợp mô hình mà không được các cá nhân chấp nhận, chẳng hạn như lập hồ sơ, sắp xếp hoặc phân loại chúng có thể ảnh hưởng đến đời sống xã hội của họ (ví dụ như dịch vụ xã hội, thẻ tín dụng).

8.4 Thiên vị

Thiên vị được định nghĩa là thiên vị đối với cái gì, người hoặc nhóm người nào hơn. Thiên vị thường phát sinh từ nhiều nguồn, bao gồm thiên vị do nhận thức của con người, thiên vị xã hội và thiên vị mang tính chất thống kê (ví dụ như thiên vị trong việc lựa chọn, thiên vị trong lấy mẫu, báo cáo thiên vị) hoặc đơn giản là các lỗi kỹ thuật. Sự thiên vị thể hiện trong các giai đoạn phát triển khác nhau của hệ thống AI và có thể ở dạng thiên vị về dữ liệu làm ảnh hưởng đến nhãn, tập dữ liệu huấn luyện, làm mất đi các tính năng/nhãn, các vấn đề về xử lý dữ liệu hoặc kiến trúc có thể ảnh hưởng đến các mô hình hoặc mô hình kết hợp. Sự thiên vị cũng có thể biểu hiện dưới dạng thiên vị được tự động hóa, tức là sự phụ thuộc quá mức vào các đề xuất của hệ thống AI. Các tác động của thiên vị về dữ liệu có thể ảnh hưởng đến

mô hình và dẫn đến kết quả không mong muốn, điều này xuất phát từ việc độ chính xác giảm cho đến việc phân loại hoàn toàn sai của các tác vụ phân loại. Loại bỏ những thiên vị này không phải lúc nào cũng khả thi và có thể tạo ra kết quả sai. Thiên vị do tập dữ liệu huấn luyện gây ra thường dựa trên các ứng dụng không chính xác hoặc không điểm xỉa đến các quy tắc và phương pháp thống kê.

Đánh giá mức độ thiên vị yêu cầu các số liệu phải được xác định và đo lường hiệu năng của hệ thống trong bối cảnh của các nhóm đối tượng cụ thể. Đã có những chỉ số cụ thể cho mục đích này. Tuy nhiên, điều cần thiết là phải hết sức thận trọng trong việc lựa chọn sử dụng các chỉ số vì những biện pháp thỏa hiệp phức tạp có thể dẫn đến kết quả không mong muốn. Các ví dụ đã có như nỗ lực điều chỉnh sự thiên vị đối với một nhóm cụ thể (các đối tượng) đã dẫn đến bối cảnh gia tăng sự thiên vị đối với một nhóm khác.

Thực tế đã tồn tại nhiều ví dụ về nguyên nhân của sự thiên vị với các biểu hiện liên quan tới hệ thống AI cũng như các biện pháp giảm thiểu tương ứng. Mô tả chi tiết của chủ đề phức tạp này được đề cập trong một báo cáo kỹ thuật chi tiết đang được xây dựng.

8.5 Tính không thể đoán trước

Khả năng dự đoán đóng một vai trò quan trọng trong khả năng chấp nhận đối với các hệ thống AI. Quan niệm rằng khả năng dự đoán tương ứng với khả năng của con người trong việc suy luận các hành động tiếp theo của hệ thống AI trong một môi trường nhất định.

Sự tin cậy vào công nghệ thường dựa trên khả năng dự đoán: một hệ thống được tin cậy nếu có thể suy luận hệ thống sẽ làm gì trong một tình huống cụ thể, ngay cả khi người ta không thể giải thích tại sao nó lại làm điều đó. Ngược lại, sự tin tưởng sẽ giảm đi nếu một hệ thống hoạt động mà không thể đoán trước được hành động của nó trong các tình huống quen thuộc.

Trong một môi trường hoạt động mà hệ thống AI tương tác với con người và nơi mà sự an toàn của con người phụ thuộc vào sự tương tác đó thì khả năng dự đoán của hệ thống là sự cần thiết chứ không phải là sự mong muốn. Sử dụng AI trong các phương tiện tự hành là một trường hợp sử dụng hiển nhiên, vì việc áp dụng rộng rãi các phương tiện tự hành dựa trên AI được dự báo là khả thi do khả năng những phương tiện đó hoạt động theo cách có thể dự đoán được. Tương tự như vậy, để chấp nhận người máy công tác có thể tương tác trực tiếp với con người thì người vận hành cần có khả năng dự đoán hành vi của người máy đó để đảm bảo an toàn cho người vận hành [55].

Trong trường hợp ô tô do con người điều khiển, cơ chế nhận thức của chúng ta đưa ra những phán đoán nhanh chóng và gần như vô thức về các hành động có thể xảy ra của con người đối với đồ vật xung quanh trên cơ sở trải nghiệm mang tính lặp đi lặp lại và đặt vào các tình huống tương tự. Ngay cả những thay đổi nhỏ trong hành vi bên ngoài cũng có thể dẫn đến mức độ khó đoán trước và trái ngược với kinh nghiệm của chúng ta. Ví dụ một chiếc ô tô tự lái có thể va chạm với một chiếc ô tô do người điều khiển vì người lái xe không có khả năng phân biệt các hành động trong tương lai của ô tô tự lái và ngược lại.

Các thuật toán học máy đưa ra những thách thức cụ thể liên quan đến khả năng dự đoán so với các kỹ

thuật lập trình truyền thống hơn. Các thuật toán ML học phân tích một lượng lớn dữ liệu và khám phá ra các kiểu mẫu và giải pháp mới. Thành phần của dữ liệu huấn luyện và các biến thể trong các mô hình cơ bản trong đó các mẫu được hiện thực hóa sẽ thiết lập các tham số đầu vào ở phạm vi mà hệ thống AI có khả năng xử lý chính xác. Các tình huống gây nhầm lẫn cho mô hình ML có thể dẫn đến hành vi không thể đoán trước, làm mất an toàn hoặc các cơ chế thay thế khác. Hơn nữa, trong trường hợp học liên tục hoặc học suốt đời, "logic" mà hệ thống ML đưa ra quyết định có thể phát triển theo thời gian và làm tăng thêm mức độ không thể đoán trước của thuật toán.

8.6 Tính không rõ ràng

Hệ thống trí tuệ nhân tạo có thể biểu hiện ở nhiều dạng không rõ ràng. Thứ nhất, bản thân mô hình AI tự nó có thể là không rõ ràng về mặt kỹ thuật, trong đó con người không dễ dàng giải thích các quá trình ra quyết định mà nó sử dụng. Thứ hai, nếu dữ liệu và nguồn dữ liệu không minh bạch, hành vi của toàn bộ hệ thống sẽ trở nên không rõ ràng đối với người quan sát bên ngoài. Thứ ba, một hệ thống AI luôn được triển khai trong bối cảnh thực thi trong một tổ chức, chẳng hạn như thu thập dữ liệu, quản lý, vận hành các kết quả AI và phát triển hệ thống. Nếu những thực tiễn này không được bộc lộ thì ngay cả một mô hình AI có khả năng diễn giải cũng trở thành một hệ thống không rõ ràng đối với người dùng và các bên liên quan khác ở bên ngoài.

Xem 9.2 để biết thêm thảo luận về các biện pháp giảm thiểu.

8.7 Những thách thức liên quan đến đặc điểm kỹ thuật của hệ thống AI

Hầu hết các lỗi của một sản phẩm đều bắt nguồn từ giai đoạn tạo lập các thông số kỹ thuật. Bởi vì giai đoạn này xác định một sản phẩm hoàn chỉnh bao gồm tính năng và môi trường hoạt động của nó, đầu ra của giai đoạn này sẽ là đầu vào của giai đoạn triển khai. Các sai hỏng của giai đoạn này có ảnh hưởng lớn tới mức khó có thể hoặc không thể sửa chữa trong các giai đoạn sau của vòng đời sản phẩm.

Sai sót trong giai đoạn này xảy ra đặc biệt khi môi trường của sản phẩm chưa được phân tích đầy đủ hoặc chi tiết. Một phân tích đầy đủ bao gồm tất cả các yếu tố môi trường có thể ảnh hưởng đến chức năng dự kiến của sản phẩm, sự quan tâm đến các mối đe dọa về an toàn và bảo mật cũng như việc thẩm tra về khuôn khổ pháp lý, quy định và đạo đức của sản phẩm. Ngoài ra, một điều quan trọng nữa là phải xem xét các khía cạnh về hiệu năng, tính năng sử dụng cho mục đích sử dụng dự kiến có tính đến bất kỳ thay đổi nào đối với môi trường triển khai cũng như các nhóm người dùng khác nhau.

Để phân tích các nguy cơ và rủi ro tiềm ẩn từ các hệ thống AI dựa trên các phương pháp học máy, điều cần thiết là phải xem xét sự cố của hệ thống do thiên vị trong dữ liệu huấn luyện hoặc do phương pháp huấn luyện sai của thuật toán. Ngoài ra, có thể tránh rủi ro bằng cách quan sát các tính năng đặc biệt của các phương pháp được sử dụng sau này và tạo dựng các tác vụ phù hợp.

Phân định rủi ro và trách nhiệm pháp lý trong hệ thống pháp luật là một nhiệm vụ phức tạp. Sử dụng AI có thể làm cho quá trình phân định đó trở nên khó khăn hơn, hoặc nó sẽ tạo ra những thay đổi về cách chúng ta đánh giá rủi ro/trách nhiệm.

Vì các hệ thống AI được sử dụng để giải quyết các tác vụ phức tạp trong môi trường không đồng nhất nên việc tạo ra một bộ đặc tính kỹ thuật đầy đủ và chính xác là điều cần thiết. Việc xác định các mục đích và mức độ có thể giải thích được (xem 9.3 để biết thêm chi tiết) là một thành phần quan trọng của bộ đặc tính kỹ thuật của bất kỳ hệ thống AI nào.

Một khi đặc tính kỹ thuật đã được xác định, nó cần phải được phổ biến cho các tác nhân riêng lẻ tham gia vào dự án. Do đó, các định nghĩa không rõ ràng về đặc điểm kỹ thuật là một nguyên nhân dẫn đến thất bại, vì điều này làm tăng nguy cơ hiểu sai và làm sai lệch các đặc tính kỹ thuật qua nhiều lần truyền đạt không chính tắc. Nói chung, các khái niệm mang tính ý tưởng được viết ra trong bản đặc tả sẽ được diễn giải bởi người tạo ra hệ thống trên cơ sở bản đặc tả đó. Điều này tạo ra sự không phù hợp đầu tiên giữa đặc điểm kỹ thuật lý tưởng và thiết kế cuối cùng. Các điểm không phù hợp khác được tạo ra khi sử dụng học máy như thuật toán cuối cùng được tạo ra bởi các khái niệm có nguồn gốc từ quá trình huấn luyện mà không có dữ liệu.

Tổng hợp các mức độ không không đảm bảo này khiến đặc tính kỹ thuật cần có các yêu cầu có thể xác minh và thẩm định để kiểm tra sản phẩm tạo ra.

8.8 Những thách thức liên quan đến triển khai các hệ thống AI

8.8.1 Thu thập và chuẩn bị dữ liệu

Trong giai đoạn thu thập dữ liệu, các nguồn dữ liệu là cần thiết để giải quyết vấn đề được xác định. Tùy thuộc vào mức độ phức tạp của vấn đề cần giải quyết, việc tìm ra bộ dữ liệu đại diện có thể là một thách thức. Trong trường hợp có dữ liệu từ nhiều nguồn thì có thể phát sinh các vấn đề về chuẩn hóa hoặc tính trọng số. Một điều quan trọng là những khiếm khuyết xử lý các nguồn dữ liệu (là đầu vào của mô hình) có thể không được phát hiện trong quá trình đánh giá mô hình, vì nó thường được tiến hành trong một môi trường cách ly thay vì trong một môi trường tích hợp.

Tại khâu này, một hoặc nhiều tập dữ liệu sử dụng để huấn luyện mô hình được kiểm tra và ghi lại mức độ phù hợp mà chúng đại diện cho dữ liệu của mô hình hoạt động.

Sau khi có được dữ liệu cần thiết, một tập các tác vụ, thường được gọi là các tác vụ chuẩn bị dữ liệu được tiến hành để làm sạch dữ liệu và đưa nó vào định dạng phù hợp cho mô hình khai thác. Một khía cạnh quan trọng trong giai đoạn này là xác minh chất lượng của dữ liệu (ví dụ dữ liệu bị thiếu, trùng lặp, dữ liệu không nhất quán hoặc dữ liệu ở định dạng sai).

8.8.2 Mô hình hóa

8.8.2.1 Yêu cầu chung

Trong giai đoạn mô hình hóa, các thuật toán đã chọn được huấn luyện để tạo ra các mô hình ứng viên. Mô hình là đại diện của một thuật toán học máy khi nó được huấn luyện với dữ liệu. Biện pháp thực hành tốt nhất là tạo lập một vài mô hình và sau đó chọn một mô hình hoạt động tốt nhất cho vấn đề cụ thể đang được nghiên cứu giải quyết. Để tạo ra một mô hình chính xác, một kỹ thuật phổ biến là tách dữ liệu có sẵn thành ba nhóm: dữ liệu huấn luyện, dữ liệu thẩm định và dữ liệu kiểm tra. Tập dữ liệu huấn luyện

là phần dữ liệu được sử dụng để huấn luyện mô hình và đảm bảo nó phù hợp. Tập dữ liệu thẩm định được sử dụng trong bước tiếp theo để thẩm định khả năng dự đoán của một mô hình được đào tạo và để điều chỉnh mô hình. Cuối cùng, tập dữ liệu kiểm tra được sử dụng trong giai đoạn thử nghiệm để cung cấp đánh giá cuối cùng về mô hình đã được đào tạo, làm cho phù hợp và sự điều chỉnh cần thiết nếu có.

Với công nghệ ngày nay, hoạt động xây dựng một hệ thống AI học máy liên quan đến các giai đoạn được mô tả trong 8.8.2.2 đến 8.8.2.5 và thường có tính lặp lại.

8.8.2.2 Thiết kế thuộc tính

Trong học máy, thuộc tính là một biến đầu vào được mô hình sử dụng để đưa ra dự đoán. Thiết kế thuộc tính là quá trình chuyển đổi dữ liệu thô thành các thuộc tính để thể hiện một cách tốt nhất các vấn đề cơ bản cần giải quyết vào trong các mô hình dự đoán, dẫn đến cải thiện độ chính xác của mô hình dựa trên dữ liệu không tường minh [56]. Các thuộc tính được tạo ra thông qua một chuỗi các bước chuyển đổi dữ liệu (ví dụ: thay đổi tỷ lệ, tùy chỉnh, chuẩn hóa, ánh xạ dữ liệu, tập hợp và tỷ lệ hóa) thường liên quan đến một số hoạt động lập trình. Cho rằng các thuộc tính là kết quả thu được từ một số bước chuyển đổi dữ liệu thì mối liên kết của nó với dữ liệu thô ban đầu có thể khó được tạo lại trừ khi quy trình chuyển đổi được ghi chép một cách cẩn thận.

Thiết kế thuộc tính ảnh hưởng mạnh đến hiệu năng của mô hình theo cách tích cực hoặc theo cách tiêu cực. Ví dụ một thuộc tính đơn nào đó góp phần chủ yếu tạo ra các dự đoán của mô hình có thể ảnh hưởng đến độ chắc chắn của mô hình, vì dự đoán cuối cùng phụ thuộc mạnh mẽ vào giá trị của biến thuộc tính đó thay vì sự phục thuộc có tính liên kết tương xứng với tất cả các thuộc tính khác. Điều này có thể dẫn đến kết quả không chính xác.

8.8.2.3 Huấn luyện mô hình

Rò rỉ mục tiêu (còn gọi là rò rỉ dữ liệu) xảy ra khi tập dữ liệu huấn luyện chứa một số thông tin liên quan đến biến được dự đoán (biến mục tiêu), là trường hợp không nên xảy ra trong giai đoạn sản xuất. Điều này có thể xảy ra, ví dụ như dữ liệu đào tạo bao gồm thông tin không có sẵn tại thời điểm dự đoán (vì biến/thuộc tính tương ứng chỉ được cập nhật sau khi giá trị mục tiêu được dự đoán). Nó cũng có thể xảy ra khi biến mục tiêu được suy ra từ dữ liệu đầu vào thông qua một biến đại diện không thể đưa vào như một thuộc tính. Các mô hình rò rỉ mục tiêu có xu hướng rất chính xác trong quá trình đánh giá nhưng lại kém hiệu quả trong hoạt động sản xuất.

Thuật toán được chọn cần được huấn luyện bằng dữ liệu huấn luyện để xây dựng mô hình. Thách thức liên quan đến giai đoạn này là xây dựng một mô hình thể hiện tính đại diện tốt hoặc phù hợp với dữ liệu huấn luyện liên quan đến vấn đề đang cần được giải quyết. Quá trình đào tạo có thể xảy ra nguy cơ tạo ra mô hình quá phù hợp hoặc thiếu sự phù hợp. Mô hình quá phù hợp là mô hình được học quá chi tiết và nó phù hợp một cách chặt chẽ với tập dữ liệu cơ sở (bao gồm cả nhiễu hoặc sai số gắn liền với tập dữ liệu). Mô hình dạng này sẽ thực thi tồi trong quá trình ra quyết định khi có các dữ liệu mới đến đầu vào. Điều này thường xuất hiện khi có quá nhiều thuộc tính được chọn là đầu vào của mô hình. Ngược

lại mô hình thiếu sự phù hợp xảy ra khi mô hình không nắm bắt được các kiểu cách cơ bản của dữ liệu, do đó các dự đoán đưa ra sẽ quá mơ hồ nên không có được các dự đoán tốt. Điều này thường xuyên xảy ra khi mô hình không có đủ các thuộc tính liên quan cần thiết. Do đó để tránh xảy ra tình trạng quá cụ thể (với quá nhiều thuộc tính) hoặc quá mơ hồ (không đủ các thuộc tính) thì điều quan trọng là phải chọn đúng các thuộc tính với lượng thông tin dự đoán phù hợp. Thiết kế một mô hình vừa vặn là một thách thức thường phát sinh trong giai đoạn huấn luyện. Tuy vậy chất lượng của các dự đoán cần được đo lường trong giai đoạn thẩm định và kiểm tra lại.

Các cách tiếp cận khác để lựa chọn và phát triển mô hình bao gồm học chuyển giao nhằm mục đích tận dụng tri thức của một tác vụ để học một tác vụ mới; hoặc là học liên kết bằng việc học các mô hình mới theo cách thức được phân bổ hoặc cộng tác.

8.8.2.4 Điều chỉnh mô hình và tối ưu hóa siêu tham số

Trong giai đoạn huấn luyện, các mô hình được hiệu chỉnh/điều chỉnh bằng cách điều chỉnh các siêu tham số của chúng. Ví dụ về siêu tham số là độ sâu của cây trong thuật toán cây quyết định, số lượng cây trong thuật toán rừng ngẫu nhiên, số cụm *k* trong thuật toán *k*-trung bình, số lớp trong mạng nơ-ron v.v.. Lựa chọn các siêu tham số không chính xác có thể là nguyên nhân dẫn đến sự thất bại của các mô hình dự đoán.

Chất lượng của một mô hình không chỉ phụ thuộc vào cấu trúc, thuật toán huấn luyện và dữ liệu của nó; một yếu tố quan trọng nữa là sự lựa chọn các siêu tham số cho mô hình. Trong một số ứng dụng, việc tối ưu hóa các siêu tham số mang lại sự cải tiến hiện đại hơn so với các thuật toán học. Tối ưu hóa siêu tham số thường tạo thành một vòng lặp bên ngoài quá trình học tập.

8.8.2.5 Đánh giá và thẩm định mô hình

Sau khi đã điều chỉnh các mô hình, chúng được đánh giá dựa trên các tập dữ liệu thẩm định để kiểm tra hiệu năng của chúng trên dữ liệu khác với tập dữ liệu được sử dụng để huấn luyện. Kỹ thuật thẩm định mô hình đơn giản thường chỉ sử dụng một tập dữ liệu thẩm định. Tuy nhiên để xây dựng các mô hình mạnh mẽ hơn có thể sử dụng kỹ thuật thẩm định chéo K-lần. Kỹ thuật này chia dữ liệu thành *k* tập con, mỗi tập con được sử dụng làm tập dữ liệu thẩm định trong khi *k*-1 tập con khác được kết hợp để tạo thành tập huấn luyện. Kết quả của *k* lần thẩm định được so sánh để xác định hiệu suất cao nhất và mô hình có độ bền vững (về độ nhạy đối với nhiễu trong dữ liệu huấn luyện). Việc lựa chọn mô hình dựa trên hiệu năng của nó so với các mô hình khác. Ví dụ về các số liệu thống kê được sử dụng để đánh giá hiệu năng của mô hình là ROC AUC (vùng dưới đường cong), ma trận nhầm lẫn (so sánh các giá trị dự đoán với các giá trị thực tế từ tập dữ liệu thử nghiệm) hoặc điểm F-1 (được tính toán dựa trên ma trận nhầm lẫn và đại diện cho điểm giao lý tưởng giữa độ chính xác và độ thu hồi).

Trong một giai đoạn kiểm tra lại riêng rẽ, mô hình đã được chọn sau giai đoạn mô hình hóa một lần nữa được kiểm tra với dữ liệu mới (tập dữ liệu kiểm tra) để có được tính nhất quán cuối cùng. Một số cài đặt cuối cùng để điều chỉnh mô hình (như ngưỡng giới hạn trong các bài toán phân loại để xác định xác suất rơi vào lớp này hay lớp kia để thỏa hiệp giữa dương tính giả và âm tính giả) được xác định cùng với

người dùng doanh nghiệp vì họ phụ thuộc vào các ứng dụng kinh doanh cụ thể.

Việc triển khai sản xuất thường diễn ra sau giai đoạn kiểm tra lại.

8.8.3 Cập nhật mô hình

Một mô hình sau khi được triển khai vào sản xuất nó có thể yêu cầu cập nhật dựa trên dữ liệu mới mà nó có được. Điều quan trọng là phải liên tục theo dõi hiệu năng/độ chính xác của mô hình để kịp thời xác định khi nào mô hình cần được huấn luyện lại/cập nhật. Cập nhật mô hình nhằm mục đích làm cho mô hình mạnh mẽ hơn và/hoặc tổng quát hóa mô hình cho các tác vụ khác nhau, hoặc để cải thiện độ chính xác của nó đối với các tập dữ liệu mới.

Một phương pháp đơn giản để cập nhật mô hình là sử dụng cả dữ liệu ban đầu và dữ liệu mới để tái huấn luyện mô hình. Có thể gặp những thách thức với cách tiếp cận này, chẳng hạn như thu thập tất cả dữ liệu vào một trung tâm, đòi hỏi khối lượng tính toán lớn để tái huấn luyện một mô hình mới dựa trên lượng dữ liệu ngày càng lớn. Một cách tiếp cận để giải quyết những thách thức này là học tăng cường, trong đó các mô hình hiện có được mở rộng dựa trên dữ liệu mới. Nói chung hiện nay việc nghiên cứu nhằm tìm ra các thuật toán hiệu quả về dữ liệu và hiệu quả về mặt tính toán để cập nhật và mở rộng các mô hình dựa trên dữ liệu mới là một lĩnh vực hiện đang được triển khai một cách tích cực.

Rủi ro chính cần lưu ý khi cập nhật một mô hình là tác động của nó đến hiệu năng. Mô hình cập nhật sẽ được thẩm định và kiểm tra lại để đảm bảo rằng không có sự suy giảm so với hiệu năng trước đó của các tác vụ ban đầu và rằng hiệu năng của bất kỳ tác vụ mới nào đều phù hợp với các ứng dụng kinh doanh cụ thể. Hơn nữa, điều quan trọng là phải đảm bảo khả năng truy nguyên và kiểm soát đối với các phiên bản khác nhau của các mô hình được triển khai trong sản xuất.

8.8.4 Lỗi phần mềm

Các phương pháp thực hiện trí tuệ nhân tạo dựa trên việc thực thi các thuật toán bằng phần mềm. Do đó, quá trình phát triển nó có chung những cạm bẫy như với mọi quá trình phát triển phần mềm khác. Các lỗi phần mềm có thể xảy ra như lỗi truy cập và điều khiển quản lý bộ nhớ, lỗi đầu vào, đầu ra; lỗi điều khiển luồng dữ liệu và lỗi dữ liệu. Các thuật toán AI thường yêu cầu một lượng tài nguyên tính toán đáng kể nên chúng thường được triển khai trên các hệ thống đa lõi. Trong những trường hợp các lỗi đồng thời xảy ra, chẳng hạn như trạng thái xung đột, tranh chấp tài nguyên hoặc hiệu ứng đo lường ("Heisenbugs") đều phải được quan tâm, xem xét.

8.9 Những thách thức liên quan đến sử dụng các hệ thống AI

8.9.1 Yếu tố tương tác người – máy (HCI)

Có nhiều cạm bẫy dựa trên các yếu tố liên quan đến con người. Có thể nhóm nhóm các yếu tố này thành bốn loại chính theo tài liệu tham khảo [57]:

1) Sử dụng, khi tự động hóa cho phép con người đạt được mục tiêu của họ;

2) Sử dụng sai, sự phụ thuộc quá mức vào tự động hóa gây ra kết quả tiêu cực không lường trước

được. Ví dụ sử dụng sai là một cá nhân quá phụ thuộc vào tự động hóa mà không chú ý đến đường đi;

3) Bỏ sử dụng, bỏ sự phụ thuộc vào tự động hóa dẫn đến kết quả tiêu cực. Ví dụ tình huống bỏ sử dụng là một cá nhân giành quyền điều khiển của hệ thống tự động hóa đang hoạt động chính xác, có thể là nguyên nhân gây ra tai nạn trên đường.

4) Lạm dụng, hệ thống tự động được thiết lập mà không tôn trọng đầy đủ lợi ích của người dùng cuối. Ví dụ lạm dụng là hệ thống thiết kế không cho phép cá nhân dễ dàng giành quyền điều khiển, kiểm soát một hệ thống tự động.

8.9.2 Áp dụng sai các hệ thống AI thể hiện hành vi thực tế của con người

Hệ thống AI có thể được thiết kế để phỏng tạo hoặc mô phỏng các đặc điểm và hành vi của con người, chẳng hạn như chữ viết tay [58], giọng nói [59], cuộc trò chuyện bằng văn bản hoặc giọng nói [60] [61]. Nếu bị các tác nhân xấu áp dụng áp dụng sai lệch, những công nghệ này có thể được sử dụng để lừa gạt các cá nhân. Có những trường hợp chatbot hoặc email-bot [62] mô phỏng con người để tạo ấn tượng cho một thành viên thực trong một dịch vụ hẹn hò.

8.10 Lỗi phần cứng hệ thống

Phần cứng hệ thống AI cần có khả năng chịu lỗi mạnh mẽ. Các lỗi phần cứng có thể là nguồn gốc vi phạm thực thi đúng của bất kỳ quá trình triển khai một thuật toán nào bằng cách làm hỏng cả luồng điều khiển và dữ liệu của nó. Đối với AI, lỗi phần cứng ảnh hưởng đến tính đúng đắn của việc thực thi thuật toán, ở cả quá trình huấn luyện và suy luận.

Bằng chứng về các lỗi phần cứng có thể bao gồm: hỏng dữ liệu, mất dữ liệu, các vấn đề về luồng dữ liệu tạm thời. Những lỗi như vậy có thể quy thành sự cố đơn lẻ hoặc sự cố kết hợp của nhiều loại lỗi khác nhau và cần được nghiên cứu thêm trong bối cảnh hiện nay của AI.

Về bản chất lỗi phần cứng có thể là vĩnh viễn (sự cố vĩnh viễn của một thành phần hoặc mô-đun trong hệ thống), tạm thời (sự cố xuất hiện tạm thời và sau đó biến mất) hoặc không liên tục (sự cố diễn ra một cách không liên tục). Các lỗi phần cứng cũng có thể là lành tính hoặc ác tính xuất phát từ các nguyên nhân ngẫu nhiên hoặc có tính hệ thống.

Các lỗi khiến một khối thiết bị ngừng hoạt động có thể là lỗi phần cứng và là lỗi lành tính do các thành phần của khối bị hỏng. Tinh vi hơn là những lỗi khiến một khối thiết bị tạo các đầu ra có vẻ phù hợp nhưng không đúng, hoặc làm cho một thành phần nào đó "hoạt động một cách có hại". Các lỗi này là lỗi mềm - là sự thay đổi trạng thái tạm thời không mong muốn của các ô nhớ hoặc các phần tử logic, nguyên nhân xuất phát thường từ các nguồn bức xạ năng lượng cao, chẳng hạn như các hạt alpha phân rã gói, gây ra bởi các nơ-tron hoặc hiệu ứng EMI bên ngoài như nhiễu điện từ, chùm tia điện từ. Nó cũng có thể gây ra bởi xuyên nhiễu nội bộ giữa các đường dẫn hoặc các linh kiện thành phần hoặc phát nhiễu độc hại, chẳng hạn như sự trục trặc của thiết bị phát xung nhịp.

Các trình điều khiển phần cứng bị lỗi cũng có thể là một khía cạnh khác để xác định lỗi phần cứng trong

máy tính do các phần mềm cài đặt cho nó bị lỗi.

Nguồn gốc và tác động của những lỗi như vậy phụ thuộc vào cả các ứng dụng AI và việc triển khai nó trên phân lớp cụ thể của một hệ thống. Các ứng dụng AI được triển khai trên các hệ thống trải rộng từ thiết bị đầu cuối chủ yếu được sử dụng để suy luận, cho đến đến tài nguyên tính toán và lưu trữ ở phạm vi đám mây sử dụng cho cả huấn luyện và suy luận. Trong vòng đời của một ứng dụng AI, mô hình được huấn luyện sẽ trải qua một số lần chuyển đổi nhằm chuyên biệt hóa các ứng dụng AI đối với nền tảng hệ thống cụ thể, ví dụ như cho thiết bị cuối hạn chế về tài nguyên.

Các mô hình lỗi khác nhau có thể lý giải các nguồn lỗi có khả năng xảy ra để từ đó có tạo dựng chiến lược chịu lỗi hiệu quả. Ví dụ các hệ thống phân tán theo thời gian thực có xu hướng kết hợp các phần cứng có sẵn (ví dụ như CPU, GPU và FPGA thông dụng), phần mềm (ví dụ như hệ điều hành thông dụng) và giao thức (ví dụ hư giao thức dựa trên ngăn xếp TPC / IP) để chạy các ứng dụng AI. Nguồn lỗi sẽ bao gồm hỏng dữ liệu, sự lặp lại thông điệp không mong muốn, chuỗi thông báo không chính xác, mất dữ liệu, độ trễ không chấp nhận được, chèn thông báo. Trong hệ thống hỗ trợ lập lịch luồng không đồng bộ đối với phần cứng như các bộ GPU, lỗi tác động tới bộ lập lịch các luồng có thể gây những ảnh hưởng lớn như không đáp ứng thời hạn thực hiện khối lượng công việc của hệ thống. Hơn nữa, việc chẩn đoán các vấn đề của bộ nhớ có thể gặp khó khăn liên quan đến các kiến trúc hệ thống cụ thể.

Các nguồn lỗi khác có thể ảnh hưởng đến hoạt động của hệ thống liên quan đến vòng đời của các ứng dụng AI. Ví dụ các nguồn lỗi bổ sung xuất hiện trong quá trình chuyển đổi mô hình - khi một mô hình được huấn luyện được tải và cài đặt xuống thiết bị đầu cuối, một hệ thống nhúng hạn chế về tài nguyên chẳng hạn, thì nó yêu cầu phải chuyên biệt hóa dữ liệu hoặc lược bớt mô hình.

9 Các biện pháp giảm thiểu

9.1 Yêu cầu chung

Biện pháp giảm thiểu là sự kiểm soát và hướng dẫn khả thi để có thể giảm thiểu những chỗ dễ tổn thương đã biết của AI được mô tả trong Điều 8. Lưu ý rằng một biện pháp kiểm soát hoặc một hướng dẫn nhất định có thể góp phần giảm thiểu một vài chỗ dễ bị tổn thương nào đó.

Một hệ thống hoạt động theo cách thức không tin cậy sẽ không được coi là có tính đáng tin cậy. Trong một số trường hợp, hệ thống có thể hoạt động bình thường nhưng tạo ra đầu ra sai lệch do dữ liệu đầu vào không chính xác mới được đưa vào. Trong bối cảnh này, sẽ có các điểm kiểm soát để xác định xem liệu sự tin tưởng còn được duy trì hay không. Các điểm kiểm soát như vậy có thể được thực hiện thường xuyên trong suốt vòng đời của hệ thống AI hoặc tại thời điểm hệ thống AI được sử dụng để ra quyết định.

9.2 Tính minh bạch

Tính minh bạch cung cấp sự tường minh đối với các thuộc tính, thành phần và quy trình của hệ thống AI. Lý tưởng nhất là một hệ thống AI được coi là minh bạch sẽ thể hiện hành vi có thể lặp lại. Tính minh bạch liên quan đến việc gia công dữ liệu, thuộc tính, mô hình, thuật toán, phương pháp huấn luyện và

quy trình đảm bảo chất lượng để kiểm tra bên ngoài. Tính minh bạch cho phép các bên liên quan đánh giá sự phát triển và vận hành của một hệ thống AI dựa trên các giá trị mà họ muốn thấy, được duy trì bởi quá trình xử lý AI. Những giá trị này có thể dựa trên các mục tiêu về sự công bằng hoặc quyền riêng tư, hoặc có thể bắt nguồn từ thế giới quan đạo đức của một bên liên quan cụ thể, chẳng hạn như đạo đức, đức hạnh hoặc hệ thống giá trị toàn cầu khác.

Tích hợp tính minh bạch vào tất cả các lớp của các quá trình AI giúp cải thiện các vấn đề gây ra bởi sự không rõ ràng được mô tả trong 8.6. Một hệ thống AI minh bạch sẽ thông báo cho các bên liên quan về địa điểm, lý do và dữ liệu nào được thu thập, đặc biệt là dữ liệu cá nhân sao cho chứng minh được siêu dữ liệu đó được thu thập tại thời điểm được phép thu thập dữ liệu. Nó cũng có thể thông báo cho các bên liên quan khi quá trình ra quyết định được tự động hóa và giải thích được quá trình đưa ra quyết định. Khi xử lý dữ liệu cá nhân dẫn đến các quyết định có hiệu lực pháp lý đối với bên liên quan, các quy định về quyền riêng tư có thể yêu cầu hệ thống AI minh bạch chấp nhận các yêu cầu can thiệp của con người trong quá trình ra quyết định có tính đến quan điểm của các bên liên quan về quá trình đó. Có một số cấp độ và thuộc tính minh bạch được xác định để phát triển các hệ thống AI khác nhau, ví dụ như trong lĩnh vực dữ liệu mở.

Việc sử dụng các ký hiệu biểu trưng, biểu tượng hoặc nhãn xếp hạng cho các hệ thống AI có thể giúp cải thiện tính minh bạch đối với nhóm các bên liên quan cụ thể. Ví dụ, Diễn đàn Kinh tế thế giới và Sáng kiến chung thế hệ AI của UNICEF [63] đề xuất các ký hiệu xếp hạng cho hệ thống AI sử dụng trong đồ chơi trẻ em mà cha mẹ/người chăm sóc có thể tiếp cận được. Khả năng giải thích của hệ thống là rất quan trọng để đạt được sự minh bạch như vậy.

Tính minh bạch của hệ thống AI liên quan đến việc gia công dữ liệu, thuộc tính, thuật toán, phương pháp huấn luyện và quy trình đảm bảo chất lượng cho các bên liên quan. Ngoài ra, trình độ kiến thức nền tảng của các bên liên quan cần phải xem như yếu tố đưa vào kế hoạch hoạt động thanh tra sao cho thuận tiện nhất. Nó có thể, nhưng không nhất thiết phải bao gồm giải thích về:

- Cách thức hoạt động của cơ chế AI đang được kiểm tra nói chung, ví dụ cách thức hoạt động của cây quyết định bằng phương pháp quy nạp;
- Loại mô hình và các tham số nào được sử dụng;
- Các biến hoặc thuộc tính cụ thể nào được sử dụng bởi mô hình; hoặc
- Cách thức một tập các biến hoặc thuộc tính ứng viên đã được chọn.

Đối với một chuyên gia được đào tạo về học máy, một lời giải thích tóm tắt ngắn gọn là đủ, chỉ rõ thủ tục lựa chọn mô hình và các biến. Đối với một người không phải là chuyên gia ở lĩnh vực nói trên thì một khóa học giới thiệu về AI và suy luận mô hình dựa vào dữ liệu sẽ rất cần thiết cùng với diễn giải hoạt động của mô hình cây quyết định, tác động của việc tham số hóa và cách thức lựa chọn các biến và thuộc tính.

Câu hỏi quan trọng về tính minh bạch là nó hoạt động như thế nào và liệu các thuật toán được sử dụng có phù hợp với mục đích đề ra hay không. Tính minh bạch làm cho dữ liệu, thuộc tính, thuật toán và

phương pháp huấn luyện trở nên có sẵn để thanh tra bên ngoài. Các biện pháp minh bạch nhằm mục đích bổ sung quyền riêng tư và lợi ích kinh doanh để nâng cao tính đáng tin cậy tổng thể của AI [64]. Trong trường hợp mô hình không rõ ràng, các phương pháp kỹ thuật có thể áp dụng để tạo ra các cấp độ minh bạch hoặc khả năng giải thích đối với các mô hình cụ thể [65].

Có thể không có sự tương ứng chặt chẽ giữa tính minh bạch và khả năng giải thích trong một hệ thống AI cũng như mức độ tin tưởng mà một bên liên quan đặt vào hệ thống đó. Tuy nhiên, tính minh bạch và khả năng giải thích cung cấp bằng chứng và thông tin quan trọng giúp các bên liên quan đưa ra nhận định về sự tin cậy của họ đối với hệ thống AI.

9.3 Khả năng giải thích

9.3.1 Yêu cầu chung

Mặc dù khả năng giải thích đơn thuần là không đủ để đảm bảo tính minh bạch của hệ thống AI, nhưng nó là một thành phần quan trọng của hệ thống AI minh bạch. Giải thích các quy trình liên quan đến phát triển, triển khai và sử dụng hệ thống AI, chẳng hạn như thực hành thu thập dữ liệu, quy trình tự kiểm soát, cam kết về giá trị và sự tham gia của các bên liên quan cũng đóng những một vai trò nhất định. Khả năng giải thích của các hệ thống AI có thể được coi là một dạng phụ của sự minh bạch của doanh nghiệp trong phạm vi trách nhiệm xã hội của doanh nghiệp.

Có thể phân loại các dạng giải thích của hệ thống AI theo: mục đích giải thích, bao gồm bối cảnh, nhu cầu của các bên liên quan; các loại hình kiến thức thu thập được và theo phương thức giải thích.

9.3.2 Mục đích giải thích

Giải thích luôn là một nỗ lực để truyền đạt sự hiểu biết. Hiệu quả giải thích có thể được cải thiện bằng cách điều chỉnh hình thức của nó cho phù hợp với ngữ cảnh đưa ra, bao gồm cả đối tượng cần giải thích cũng như mức độ hiểu biết để truyền đạt cho phù hợp [66].

Nỗ lực giải thích có thể đưa ra nhiều cách thức giải thích khác nhau nhưng đều phải hợp lệ, tùy thuộc vào việc các bên liên quan tìm kiếm để:

- Hiểu về cách thức làm thế nào để có được kết quả;
- Lĩnh hội các tri thức có trong kết quả; hoặc
- Hiểu rõ các sở cứ dựa vào đó kết quả đưa ra là hợp lệ.

Đối tượng giải thích có thể bao gồm bản thân hệ thống AI và kết quả do hệ thống tạo ra.

Tính giải thích được của các hệ thống AI nhằm mục đích cung cấp sự hiểu biết về các quy trình góp phần tạo nên sự trung thực, chính xác và hợp lý cho các kết quả của nó để bác bỏ những quan sát mang tính quy kết rằng hệ thống hoạt động chỉ là hình thức. Việc hiểu rõ các giải thích từ các bên liên quan có thể được hỗ trợ bằng việc tuân thủ các hướng dẫn và tiêu chuẩn phù hợp.

Các giải thích liên quan đến hệ thống AI cũng có thể cung cấp những lý do chính đáng về tính hợp lệ, sự phù hợp và tính hợp pháp của các kết quả cũng như các quyết định, hành động dựa trên các kết quả

đó. Những giải thích như vậy làm cho một hệ thống AI dễ giám sát và dễ thảo luận hơn, nhất là đối với các bên liên quan bị ảnh hưởng bởi các quyết định và hành động mà hệ thống đưa ra.

Giải thích không mang tính tuyệt đối nhưng cần được xác định rõ những vấn đề liên quan đến mô hình đích và người nhận sự giải thích. Giải thích phải hiểu được và làm rõ các vấn đề. Thông tin cần được thể hiện theo cách thức nào đó để tăng cường tính trung thực đối với cách nhìn nhận của con người đối với hệ thống và đối với chính bản thân hệ thống đó [66][71].

9.3.3 Giải thích trước và giải thích trước sau

Giải thích trước là giải thích các thuộc tính và tính năng chung của một hệ thống trước khi sử dụng hệ thống đó. Hệ thống AI được giải thích từ trước bằng việc cung cấp thông tin cho các bên liên quan ngoại trừ nhà phát triển, về các thuộc tính và tính năng của một hệ thống, trước khi sử dụng hệ thống đó.

Giải thích sau là giải thích các thuộc tính và tính năng của hệ thống có vai trò trong việc đưa ra quyết định. Sử dụng các biểu trưng có thể tăng cường khả năng giải thích và dẫn đến nâng cao tính đáng tin cậy của AI.

Các giải thích trước và trước sau phục vụ các chức năng khác nhau. Giải thích trước cố gắng tạo lập niềm tin rằng hệ thống được thiết kế tốt và phục vụ mục đích sử dụng đề ra. Nó nhằm mục đích tạo lập niềm tin với người dùng và thúc đẩy việc sử dụng hệ thống AI đó ngay từ đầu. Giải thích sau cho phép giải thích các kết quả của thuật toán cụ thể và hoàn cảnh mà chúng được thực hiện. Điều này có nghĩa là giải thích trước được coi là quan trọng để tạo lập niềm tin vào hệ thống AI, nhưng sẽ không có được tính minh bạch của hệ thống nếu như không tiếp cận đến hoạt động giải thích sau của hệ thống đó.

Lý tưởng nhất là một hệ thống AI có sự nhất quán giữa các giải thích trước và giải thích sau. Các thuộc tính và tính năng đưa ra trong giải thích trước đó sẽ được chứng minh bởi thông tin được bộc lộ thông qua giải thích sau thể hiện hoạt động của các thuật toán cụ thể.

Mục 9.3.4 đề cập đến khả năng giải thích sau.

9.3.4 Các cách tiếp cận để giải thích

Có thể phân loại các cách tiếp cận để giải thích dựa trên giai đoạn, phạm vi và mức độ chi tiết của các giải thích được tạo ra. Các giải thích có thể được tạo ra trong các giai đoạn khác nhau của quá trình phát triển mô hình AI:

1) Trước mô hình hóa;
2) Trong khi mô hình hóa; và
3) Sau mô hình hóa.

Giai đoạn trước mô hình hóa phục vụ việc hiểu dữ liệu trước khi xây dựng mô hình. Có một nhóm các phương pháp có thể được thực hiện để hiểu một tập dữ liệu nào đó nhằm cung cấp thông tin về sự phát triển tiếp theo của các mô hình AI (ví dụ: phân tích các mặt của dữ liệu, trực quan hóa dữ liệu, chuẩn hóa dữ liệu, phân tích bằng phương pháp toán học các tập dữ liệu) [72] - [76].

Giai đoạn mô hình hóa phục vụ việc phát triển các mô hình AI có thể giải thích các quyết định của chúng hoặc vốn dĩ đã giải thích được [77] - [79].

Giai đoạn sau mô hình hóa phục vụ việc tạo ra các giải thích cho các quyết định cho một mô hình AI không diễn giải được [80] - [88].

Quá trình giải thích một mô hình AI có thể được mô tả như sau:

- Cục bộ, bằng việc giải thích quá trình mô hình ra quyết định đối với một cặp đầu vào / đầu ra nhất định; hoặc
- Toàn cục, bằng việc giải thích mang tính logic về một khái niệm chung hoặc một loại kiểu cách dữ liệu bên trong một mô hình.

Ngoài ra, các giải thích có thể được tạo ra với các mức độ chi tiết khác nhau. Với một mạng nơ-ron sâu, một cấp độ giải thích nào đó có thể đề cập thảo luận về vai trò của từng lớp trong kết quả dự đoán. Có thể hiểu chi tiết hơn bằng việc kiểm tra vai trò của từng nơ-ron trong một lớp nhất định [89] - [95].

9.3.5 Các phương thức giải thích sau

9.3.5.1 Yêu cầu chung

Các phương thức giải thích có thể được phân loại theo truy nguyên, nhận thức và biện luận.

Ba phương thức giải thích này có thể khác nhau, vì một tổ chức có thể đưa ra lời giải thích theo truy nguyên mà không cần đưa ra lời giải thích mang tính nhận thức hoặc tính biện luận. Ví dụ nêu bật đặc điểm giới tính trong kết quả của thuật toán quyết định cho vay tín dụng như là một phần của sự giải thích nguyên nhân mà thuật toán đưa ra quyết định đó, nhưng nó không trả lời các câu hỏi về vai trò chức năng của giới tính trong khả năng vay tín dụng của một người, hoặc theo những tiêu chuẩn nào thì nó cho là hợp lệ để biện minh cho quyết định vay tín dụng trên cơ sở giới tính đó.

Do đó giải thích mang tính đầy đủ của hệ thống AI có thể bao gồm các thuộc tính sau:

- Chuỗi các truy nguyên theo dõi cách thuật toán đưa ra quyết định;
- Vai trò chức năng của các thuộc tính được đánh giá đối với các hiện tượng được mô hình hóa;
- Các nguyên tắc và tiêu chuẩn về đạo đức cùng các quy định khác để minh chứng kết quả đầu ra của các thuật toán.

Việc lựa chọn các thuộc tính giải thích như trên phụ thuộc vào việc giải thích đó dành cho ai, mục đích của giải thích là gì và mức độ tin tưởng mong muốn đạt được đối với ứng dụng.

9.3.5.2 Giải thích theo truy nguyên: cái gì hoạt động như thế nào

Đối với mục tiêu để hiểu cách thức một hệ thống AI có được các kết quả thì giải thích truy nguyên bao gồm một chuỗi các truy nguyên để giải thích các cơ chế, trong đó các thuộc tính đầu vào được xử lý để tạo ra kết quả nhất định.

Kết quả của việc theo dõi chuỗi truy nguyên trong quá trình học máy phụ thuộc vào mức độ trừu tượng được chọn. Đó là các thuộc tính định tính (ví dụ hình dạng của đối tượng), phép tính toán ẩn dụ (ví dụ

giá trị véc-tơ), các vấn đề vật lý thực tế (ví dụ trạng thái nạp của thanh ghi bộ xử lý). Chúng đều có những vai trò nhất định trong lịch sử truy nguyên của một quá trình AI, giải thích truy nguyên để hiểu cách thức một kết quả được tạo ra và có thể tiến hành ở bất kỳ cấp độ nào [96].

Mức độ trừu tượng nào được coi là hữu ích tùy thuộc vào từng mục tiêu giải thích. Với một hệ thống AI có thể diễn giải, mức độ trừu tượng cao nhất của yếu tố quyết định nào đó mà hệ thống sử dụng, cùng với trọng số của chúng đối với kết quả cuối cùng có thể được làm cho hài hòa với các thuộc tính mang tính định lượng và có ý nghĩa đối với con người. Hơn nữa, giải thích truy nguyên có thể hỗ trợ khi xuất hiện các biện pháp can thiệp phản thực tế [97]. Nghĩa là nó có thể mang lại sự hiểu biết về kết quả được tạo ra sẽ cần thay đổi như thế nào, liệu các thuộc tính đầu vào đã được sửa đổi chưa.

9.3.5.3 Giải thích theo nhận thức: Làm thế nào chúng ta biết nó hoạt động

Đối với mục tiêu biện luận theo nhận thức, nghĩa là giải thích tại sao một kết quả được tạo ra bởi thuật toán là đúng, một giải thích được coi là thành công nếu nó diễn giải được các mối quan hệ chức năng hoặc logic trong các hiện tượng được mô hình hóa. Có nghĩa là nó không phải là mô tả liên quan đến bản thân hệ thống, mà nó là mô tả liên quan đến các đặc điểm của thế giới mà hệ thống hướng đến.

9.3.5.4 Giải thích theo biện luận: Nó hoạt động dựa trên cơ sở nào

Đối với một quyết định tự động từ hệ thống AI, đó là sự giải thích cho tính hợp lệ của kết quả. Điều này vượt ra ngoài phạm vi của giải thích theo truy nguyên và giải thích theo chức năng trong một bối cảnh xã hội nào đó để truyền đạt các nguyên tắc, sự kiện và tiêu chuẩn làm sở cứ đưa ra quyết định. Cách thức giải thích này cho thấy lý do tại sao kết quả đưa ra là phù hợp, công bằng, hợp lệ dựa trên tình huống hiện tại.

Giải thích theo biện luận có thể đề cập đến các thuộc tính của hệ thống AI như thuật toán, dữ liệu được sử dụng, các thuộc tính quyết định, nhưng sẽ không được coi là đầy đủ nếu không tham chiếu đến các thực tế về thể chế, xã hội và việc triển khai hệ thống. Điều này bao gồm các quy định, tiêu chuẩn và quy trình tổ chức phù hợp với trường hợp sử dụng.

Chức năng giải thích theo biện luận được coi là thành công nếu lập luận hỗ trợ kết quả được tạo ra mang tính hệ thống. Do đó giải thích thành công mở ra hướng thuận lợi để xem xét và thảo luận một cách kỹ lưỡng các kết quả, kết quả của hệ thống có thể được đánh giá lại dựa trên các lập luận phản bác để từ đó có sự đảo ngược hoặc điều chỉnh quyết định một cách phù hợp.

9.3.6 Cấp độ giải thích

Cấp độ giải thích phù hợp của một hệ thống AI có thể được chọn cho bối cảnh của trường hợp sử dụng mà hệ thống đang áp dụng. Do đó các yêu cầu được xác định rõ ràng đối với các cấp độ giải thích khác nhau có thể hỗ trợ lựa chọn loại hình AI cho một ứng dụng dựa trên năng lực giải thích theo cấp độ của nó. Ví dụ các hệ thống không có khả năng giải thích sẽ không cung cấp các giải thích theo truy nguyên có ý nghĩa về hoạt động của chúng sẽ không thích hợp để sử dụng trong các sản phẩm hoặc dịch vụ mong muốn có cấp độ giải thích cao. Cấp độ có thể giải thích nào đó được áp dụng sẽ được đánh giá

theo từng trường hợp cụ thể.

Khi lựa chọn cấp độ có thể giải thích cần thiết cho một hệ thống AI, các quan tâm đối với ứng dụng có thể bao gồm những vấn đề:

- Hệ thống AI sử dụng dữ liệu nhạy cảm của các cá nhân làm dữ liệu đầu vào;
- Kết quả của hệ thống AI được sử dụng theo cách thức có sự tác động đáng kể đến lợi ích của các cá nhân;
- Hậu quả của việc ra quyết định không đúng dựa trên AI là tương đối quan trọng;
- Ứng dụng có thể dẫn hạn chế quyền tự chủ của người dùng hoặc của các bên thứ ba;
- Hệ thống có tác động không nhỏ đối với những người ngoài cuộc và cộng đồng xã hội lớn hơn so với phạm vi mà nó được triển khai, ví dụ như chỉ hiển thị một số quảng cáo việc làm cho nam giới dẫn đến gia tăng bất bình đẳng giới.

Cấp độ giải thích có thể khác nhau dựa trên nhu cầu cụ thể của các nhóm bên liên quan khác nhau, các khía cạnh dữ liệu mà dựa vào đó hệ thống AI đưa ra kết quả, nhu cầu cần có sự can thiệp của con người đối với các quyết định dựa trên kết quả AI, nhu cầu để các bên liên quan bày tỏ quan điểm riêng cũng như những thách thức gặp phải đối với các quyết định đó.

9.3.7 Đánh giá các giải thích

Đánh giá chất lượng giải thích cũng là điều quan trọng. Nó bao gồm việc xem xét ở các khía cạnh sau:

- Tính liên tục, giải thích liên quan đến các dự đoán có điểm lân cận nhau là gần như tương đương;
- Tính nhất quán, nếu chúng ta thay đổi mô hình sao cho sự tham gia của một tính năng nhất định vào kết quả dự đoán tăng lên, thì điểm số quan trọng của thuộc tính đó ước lượng bằng phương pháp giải thích sẽ không bị giảm;
- Tính chọn lọc, các giải thích dựa trên mức độ quan trọng, với mong muốn rằng sự tham gia, được phân bổ theo các thuộc tính sẽ có tác động mạnh nhất đến dự đoán được tạo ra nếu sự tham gia đó là quan trọng nhất. Nghĩa là việc loại bỏ một tính năng (hoặc một tập hợp các tính năng) có điểm số liên quan cao nhất sẽ dẫn đến sự thay đổi mạnh mẽ trong kết quả đầu ra của mô hình. Điều này đảm bảo rằng các thuộc tính đúng được phân biệt bằng những nội dung liên quan trong phần giải thích được tạo ra.

Ngoài ra cũng cần xem xét sự trả giá giữa tính chính xác và tính dễ hiểu của sự giải thích [98] - [105].

9.4 Khả năng điều khiển

9.4.1 Yêu cầu chung

Có thể đạt được khả năng điều khiển bằng cách cung cấp các cơ chế đáng tin cậy để người vận hành tiếp quản quyền kiểm soát từ hệ thống AI. Để đạt được khả năng điều khiển, các câu hỏi cần giải quyết trước tiên bao gồm ai được cấp quyền điều khiển cái gì, hệ thống AI thuộc sở hữu của ai, trong bối cảnh hệ thống đó có nhiều bên liên quan tham gia, ví dụ như nhà cung cấp dịch vụ hoặc nhà cung cấp sản

phẩm, nhà cung cấp các thành phần cấu thành AI, người dùng hoặc tác nhân có thẩm quyền quản lý.

Những mô tả dưới đây thể hiện sự cần thiết phải tích hợp các điểm kiểm soát trong vòng đời của hệ thống AI như một khâu để đưa ra quyết định đáng tin cậy.

9.4.2 Các điểm điều khiển bằng con người trong vòng lặp

Xét về vai trò của con người trong vòng đời các hệ thống AI, có hai vai trò liên quan đặc biệt được coi như các điểm điều khiển bằng con người trong vòng lặp:

- Những người ra quyết định có quyền tự quyết và quyền tự chủ trong quá trình ra quyết định cuối cùng có cân nhắc đến kết quả đưa ra bởi hệ thống AI để tăng cường khả năng ra quyết định của bởi con người;
- Các chuyên gia trong từng lĩnh vực phù hợp cung cấp các phản hồi để không chỉ đánh giá lại mức độ tin tưởng vào hệ thống mà còn cải thiện hoạt động của hệ thống. Trong bối cảnh này, kết quả được kiểm tra/ngữ cảnh hóa bởi các chuyên gia trong từng lĩnh vực phù hợp và được coi là quan trọng đối với hệ thống AI. Vì các chuyên gia đó có thể chỉ ra các tương quan giả hoặc lý do tại sao một hệ thống hoạt động vẫn có thể hoạt động theo một cách thức nào đó trong khi dữ liệu không khả dụng cho hệ thống AI.

9.5 Các chiến lược giảm tính thiên vị

Hiện tồn tại nhiều chiến lược giải quyết sự thiên vị:

- Xem xét các yêu cầu pháp lý và các yêu cầu khác liên quan đến tính thiên vị và có thể xác định một cách tường minh trong giai đoạn xác định các yêu cầu hệ thống, bao gồm cả việc thiết lập các ngưỡng thích hợp;
- Phân tích xuất xứ và tính đầy đủ của nguồn dữ liệu để có thể bộc lộ những rủi ro, soát xét các quy trình được sử dụng để thu thập và chú giải dữ liệu;
- Các kỹ thuật thuần túy có thể sử dụng như một phần của quá trình huấn luyện mô hình để phát hiện và giảm thiểu tính thiên vị;
- Các kỹ thuật kiểm tra và đánh giá cụ thể sử dụng để phát hiện tính thiên vị;
- Các thử nghiệm hoặc đánh giá hoạt động thường xuyên sử dụng để phát hiện các vấn đề liên quan đến tính thiên vị trong bối cảnh sử dụng thực tế.

Mỗi cách giải quyết trên có những ưu và nhược điểm của nó. Điều tra rủi ro liên quan đến tính thiên vị và tài liệu hóa các kỹ thuật giảm thiểu sẽ giúp việc tạo lập tính đáng tin cậy đối với AI.

9.6 Quyền riêng tư

Các phương pháp cú pháp (chẳng hạn như k-nặc danh) hoặc các phương pháp ngữ nghĩa (chẳng hạn như quyền riêng tư khác biệt) được sử dụng để khử danh tính dữ liệu cá nhân [52]. Ngay cả khi dữ liệu được khử nhận dạng, thì khi có sẵn các dữ liệu từ nhiều nguồn AI vẫn có thể nhận dạng lại dữ liệu bằng cách sử dụng suy luận dựa trên dữ liệu từ các nguồn khác đó. Ví dụ nghiên cứu [53][54] cho thấy giải

pháp k-nặc danh có thể vẫn chưa để để thỏa mãn.

Bất kể phương pháp khử danh tính ban đầu là gì đều có thể phải quản lý rủi ro về tái nhận dạng danh tính bằng các thỏa thuận sử dụng dữ liệu giữa các bên khi thu nhận dữ liệu.

9.7 Độ bền vững, khả năng phục hồi và độ bền vững

Công bố [30] chỉ ra "khả năng" của hệ thống là một trong những thành phần quan trọng để đạt được tính đáng tin cậy. Khả năng có thể được mô tả như một đặc tính của hệ thống để thực hiện một tác vụ cụ thể, có thể được đánh giá theo một số thuộc tính bao gồm tính tin cậy, khả năng phục hồi và độ bền vững.

Tính tin cậy là khả năng của một hệ thống hoặc một thực thể trong hệ thống đó thực hiện các chức năng cần thiết của nó ở các điều kiện xác định trong một khoảng thời gian cụ thể [106]. Nói cách khác, một hệ thống AI đáng tin cậy tạo ra các đầu ra giống nhau cho các đầu vào giống nhau một cách nhất quán.

Hệ thống AI cũng như các loại hình hệ thống phần mềm khác, lỗi phần cứng có thể ảnh hưởng đến việc thực thi một cách đúng đắn của thuật toán. Khả năng chịu lỗi là khả năng hệ thống tiếp tục hoạt động khi xuất hiện sự gián đoạn, lỗi hoặc hỏng hóc trong hệ thống, nhưng có khả năng làm suy giảm năng lực của hệ thống đó. Xét về tổng thể, một hệ thống hoặc thiết bị hoạt động chính xác, đáp ứng các yếu tố đầu vào của nó thì hệ thống đó được xem như an toàn trong hoạt động [107].

Khả năng phục hồi là khả năng hệ thống có thể phục hồi tình trạng hoạt động một cách nhanh chóng sau khi xảy ra sự cố. Khả năng phục hồi liên quan đến tính tin cậy nhưng có cấp độ dịch vụ mong muốn và được kỳ vọng là khác nhau. Kỳ vọng về khả năng phục hồi có thể thấp hơn theo cách xác định của các bên liên quan, điều này cũng đúng với đề xuất về khả năng khôi phục (xem tài liệu tham khảo [42], mục 11.5).

Đối với các hệ thống AI, độ bền vững thường sử dụng để mô tả khả năng sau cùng của một hệ thống duy trì mức hiệu năng của nó trong bất kỳ trường hợp nào, bao gồm cả sự can thiệp từ bên ngoài hoặc các điều kiện môi trường khắc nghiệt. Độ bền vững bao gồm khả năng phục hồi, tính tin cậy và nhiều thuộc tính tiềm năng khác nữa liên quan đến hoạt động đúng đắn của một hệ thống theo dự định của các nhà phát triển. Rõ ràng là hoạt động hoàn hảo của một hệ thống có liên quan trực tiếp hoặc ảnh hưởng tới sự an toàn của các bên liên quan trong một môi trường/bối cảnh nhất định. Ví dụ một hệ thống AI dựa trên ML mạnh mẽ sẽ có khả năng khái quát hóa được các đầu vào chưa biết, ví dụ như làm mất đi sự quá phù hợp của mô hình. Để đạt được điều đó thì điều cần thiết là phải huấn luyện các mô hình một cách thực chất, hoặc các mô hình sử dụng bộ dữ liệu huấn luyện lớn, bao gồm cả dữ liệu huấn luyện có nhiễu.

9.8 Giảm thiểu lỗi phần cứng hệ thống

Hệ thống có độ bền vững và khả năng chịu lỗi đạt được bằng các phương pháp khác nhau liên quan đến kiến trúc và thiết kế chi tiết của phần cứng cũng như toàn bộ quá trình phát triển của nó. Do đó mọi giai đoạn trong chu kỳ sống của sản phẩm, đặc biệt là giai đoạn thiết kế và xây dựng đặc tính kỹ thuật

của hệ thống đều nằm trong phạm vi đề cập ở mục này.

Một trong những phương pháp này là khai thác sự tính dự phòng để che giấu hoặc khắc phục các lỗi hỏng hóc và do đó duy trì được cấp độ hoạt động mong muốn. Các lỗi phần cứng có thể được xử lý bằng sử dụng dự phòng bằng phần cứng (ví dụ: n-lớp tại tầng nào đó hoặc có cấp độ dự phòng phù hợp), bằng thông tin (ví dụ như các bit kiểm tra) hoặc bằng thời gian (ví dụ tính toán lại vào các thời điểm khác nhau, thường là ngẫu nhiên). Các lỗi phần mềm được bảo vệ bằng dự phòng phần mềm (ví dụ đa dạng hóa phần mềm hoặc các hình thức giảm thiểu khác).

Loại lỗi trước đây được giảm thiểu bằng cách kết hợp phần cứng bổ sung vào thiết kế để phát hiện hoặc thay thế ảnh hưởng của thành phần bị lỗi. Dự phòng phần cứng có thể là ở chế độ tĩnh hoặc động; nó có thể bao gồm từ một bản sao đơn giản đến các cấu trúc phức tạp, chuyển đổi các thiết bị dự phòng khi các thiết bị đang hoạt động bị lỗi.

Để tránh các tác động xấu của các lỗi với nguyên nhân thông thường, chẳng hạn như ảnh hưởng từ điều kiện môi trường hoặc điểm yếu của các công nghệ cảm biến cụ thể thì cần có nhiều biện pháp hơn (ví dụ: sử dụng tính đa dạng). Ngoài ra, các biện pháp chẩn đoán khác nhau có thể giúp phát hiện lỗi trong thời gian hoạt động và thực hiện các biện pháp đối phó hoặc chuyển hệ thống sang trạng thái an toàn.

Có thể tìm thấy mô tả toàn diện về các phương pháp và quy trình để triển khai phần cứng an toàn khi có hỏng hóc, mô tả về các cấp độ hoạt động an toàn có thể chứng nhận được ở trong IEC 61508 [107].

9.9 Tính an toàn trong hoạt động

Các chức năng cụ thể ở nhiều khía cạnh liên quan được thực thi để đảm bảo an toàn cho hoạt động của hệ thống. Các chức năng này có thể là một phần không thể thiếu của chức năng điều khiển của một hệ thống hoặc một hệ thống chuyên dụng có giao tiếp với các hệ thống đang được xem xét. Ví dụ, đối với các hệ thống AI, chức năng liên quan đến an toàn có thể giám sát các quyết định do AI thực hiện để đảm bảo rằng chúng nằm trong phạm vi có thể chấp nhận được hoặc đưa hệ thống vào trạng thái xác định trong trường hợp chúng phát hiện ra hành vi có vấn đề.

IEC 61508 [107] đưa ra phương pháp tiếp cận chung cho tất cả các hoạt động trong vòng đời an toàn của các hệ thống bao, gồm các phần tử điện và/hoặc điện tử, và/hoặc các phần tử điện tử có thể lập trình được sử dụng để thực hiện các chức năng an toàn. Nó là cơ sở cho các tiêu chuẩn quốc tế trong lĩnh vực sản phẩm và ứng dụng để giải quyết các vấn đề liên quan đến tính an toàn của hệ thống. ISO 26262 [108], IEC 62279 [109] và IEC 61511 [110] là các ví dụ về các tiêu chuẩn thích ứng theo lĩnh vực cụ thể cho ngành ô tô, đường sắt và công nghiệp chế biến. IEC 61508 [107] có thể áp dụng cho tất cả các hệ thống liên quan đến an toàn điện và/hoặc điện tử, và/hoặc phần tử điện tử có thể lập trình (E / E / PE) thuộc bất kể ứng dụng là gì. Nó chủ yếu quan tâm đến những hệ thống mà sự cố của nó có thể ảnh hưởng đến sự an toàn của con người và/hoặc môi trường. Tuy nhiên, người ta cũng cho rằng hậu quả của sự hỏng hóc có thể gây ra những ảnh hưởng nghiêm trọng về kinh tế. Trong các trường hợp như vậy, các tài liệu nói trên có thể được sử dụng để chỉ rõ hệ thống E / E / PE cụ thể được sử dụng để bảo vệ thiết bị hoặc sản phẩm.

9.10 Kiểm tra và đánh giá

9.10.1 Yêu cầu chung

Có nhiều cách tiếp cận khác nhau để kiểm tra và đánh giá các hệ thống AI. Mặc dù khả năng ứng dụng và hiệu quả của chúng có thể khác nhau tùy từng trường hợp, nhưng thông thường sẽ cần sự kết hợp của nhiều phương pháp để đạt được mức độ đáng tin cậy có thể chấp nhận được.

9.10.2 Phương pháp thẩm định và xác minh phần mềm

9.10.2.1 Yêu cầu chung

Để đạt được tính đáng tin cậy, các hệ thống phần mềm truyền thống (không phải AI) dựa trên hai trục sau:

- Một kiến trúc cho phép dự phòng hoặc giám sát các chức năng quan trọng; và
- Thẩm định và xác minh mã nguồn, bao gồm việc chứng minh bằng kỹ thuật rằng mã thực thi đáp ứng yêu cầu và được kiểm tra đầy đủ. Hiện tại việc chứng minh dựa trên thực tế là hành vi của hệ thống đã được biết trước và có thể xác định được.

Theo tài liệu tham khảo [111], thẩm định là "xác nhận, thông qua việc cung cấp bằng chứng khách quan, rằng các yêu cầu cho một mục đích sử dụng hoặc ứng dụng cụ thể đã được đáp ứng. Chú thích 1: hệ thống đúng đã được tạo lập". Xác minh là "xác nhận, thông qua việc cung cấp bằng chứng khách quan, rằng các yêu cầu cụ thể đã được đáp ứng. Chú thích 1: hệ thống đã được tạo lập đúng" [24].

Các hệ thống phần mềm phải tuân theo các phương pháp thẩm định, xác minh và kiểm tra phần mềm một cách chính thức, chẳng hạn như được quy định trong tài liệu tham khảo [111], trong đó có nêu mục tiêu chính của kiểm thử phần mềm.

"Cung cấp thông tin về chất lượng của hạng mục kiểm thử và bất kỳ rủi ro nào có thể tồn tại liên quan đến hạng mục được kiểm thử để tìm các khuyết tật trong hạng mục kiểm thử trước khi sử dụng; và để giảm thiểu rủi ro cho các bên liên quan do chất lượng sản phẩm kém".

Theo thiết kế, yếu tố tất định của các hệ thống AI thường kém hơn so với các hệ thống phần mềm truyền thống và hiếm khi được giải thích một cách cặn kẽ. Phần mềm của hệ thống AI bao gồm cả các thành phần AI và không phải AI.

Mặc dù tất cả các thành phần của hệ thống AI cần tuân theo phần mềm và phần cứng thực tế được chấp nhận (bao gồm các bài kiểm tra cho các khối và chức năng) cho việc hoạt động chính xác, nhưng các thành phần AI của nó sẽ sử dụng phiên bản sửa đổi thực tế như thảo luận ở dưới đây.

Đối với hệ thống AI, cần có các kiểm thử chức năng để có thể xử lý tính không chắc chắn khi áp dụng. Đó là một thách thức đối với việc xác định và kiểm thử các yêu cầu đối với các thành phần phần mềm không có yếu tố mang tính tất định bằng các tiêu chuẩn và thông lệ hiện có. Đây được gọi là "vấn đề tiên đoán", mô tả những khó khăn trong việc xác định liệu một phép kiểm thử riêng biệt nào đó có đáp ứng các tiêu chí thành công trong việc kiểm thử hay không. Sự phổ biến của vấn đề này trong các hệ thống

AI đòi hỏi các nỗ lực trong hoạt động tiêu chuẩn hóa để thúc đẩy việc tạo lập các kỹ thuật thẩm định và xác minh mới.

9.10.2.2 Phương pháp chính thức

Có thể sử dụng các phương pháp chính thức để kiểm tra và đánh giá mạng nơ-ron nhân tạo cho mục đích thẩm định và xác minh phần mềm. Để làm được như vậy, một số chỉ số có thể được sử dụng, chẳng hạn như:

- Tính không chắc chắn, tương quan với sự thay đổi đáp ứng của mạng nơ-ron để kiểm tra xem liệu sự tổng quát hóa của nó có tạo ra hành vi không ổn định hay không;
- Không gian ổn định tối đa, tương quan với khả năng của hệ thống AI chứng minh rằng việc phân loại được thực hiện sẽ ổn định xung quanh tập huấn luyện.

9.10.2.3 Kiểm tra thực nghiệm

Có nhiều kỹ thuật khác nhau để kiểm tra thực nghiệm các giải pháp không có tính tất định cho mục đích thẩm định và xác minh phần mềm, bao gồm:

- Thử nghiệm biến hóa – một kỹ thuật thiết lập mối quan hệ giữa đầu vào và đầu ra của hệ thống và dựa vào việc chạy nhiều lần thử nghiệm và so sánh kết quả. Nó thường được sử dụng trên các hệ thống có "vấn đề tiên đoán" [112];
- Hội đồng chuyên gia – cho các hệ thống AI được xây dựng để thay thế cho phán đoán của các chuyên gia, hội đồng được thành lập để xem xét kết quả thử nghiệm. Cách tiếp cận này có thể nảy sinh ra thách thức mới do việc sự không đồng ý của các chuyên gia về kết quả đưa ra [113];
- Chấm điểm – Kỹ thuật đo lường hiệu năng của một hệ thống trên các tập dữ liệu được thiết kế, chuẩn bị kỹ càng, công khai và khả dụng để kiểm tra so sánh, đối soát giữa các hệ thống khác nhau [114]. Trong phương pháp của AI về nhận dạng mẫu và các ứng dụng tương tự thì kiểm tra chấm điểm là cách thực hành để tạo lập sự tin tưởng đối với một phương pháp nhất định [115].

9.10.2.4 So sánh thông minh

Khi không có phương pháp đánh giá tự động nào, việc so sánh khả năng thông minh của hệ thống AI và con người có thể cung cấp sự tin tưởng về chất lượng hệ thống AI bằng cách khẳng định các chức năng thực hiện của hệ thống AI. Cách tiếp cận này dựa trên việc so sánh các chỉ số nhất định với ngưỡng theo tiêu chí đã cho. Sau đó có thể áp dụng một vài phương pháp khác nhau (ví dụ như hệ số hòa hợp, thực nghiệm Pearson) với các môi trường khác nhau (ví dụ "hộp cát" hoặc ngăn vật lý).

9.10.2.5 Thử nghiệm trong môi trường mô phỏng

Trong một số trường hợp khi tác vụ thực hiện bởi hệ thống AI đặc trưng bởi hành động vật lý đối với môi trường (ví dụ đối với AI được nhúng trong người máy), việc đánh giá hiệu năng và phân tích sự tuân thủ với các yêu cầu liên quan đến rủi ro cần phải được thực hiện trong thực tế hoặc môi trường có thể đại diện cho thực tế. Để xác định phạm vi hoạt động của AI nhúng, cần thực hiện các thử nghiệm trong môi trường được kiểm soát để thúc đẩy khả năng chấp nhận đối với các hệ thống cơ điện tử thông minh. Có

thể thực hiện thử nghiệm vật lý trong buồng khí hậu, thử nghiệm rung, sốc, gia tốc không đổi để đánh giá hiệu năng của hệ thống trong các điều kiện khắc nghiệt và để xác định chính xác các điều kiện biên cho vận hành. Việc đánh giá các hệ thống AI trong môi trường mở và thay đổi có thể gặp phải rất nhiều các trường hợp thử nghiệm khác nhau, phát triển các môi trường thử nghiệm ảo để đánh giá bằng mô phong mô phỏng cũng có thể là cách tiếp cận hữu ích.

9.10.2.6 Thử nghiệm hiện trường

Do sự khác biệt giữa môi trường thử nghiệm và điều kiện vận hành thực tế, nên thử nghiệm hiện trường thường là một cách rất hiệu quả để cải thiện chất lượng của hệ thống được triển khai bằng việc thử nghiệm hiệu năng, tính hiệu quả hoặc độ bền của hệ thống.

Một số ví dụ và lĩnh vực nổi bật áp dụng loại hình thử nghiệm này là:

- Thử nghiệm nhận dạng khuôn mặt [116];
- Thử nghiệm hệ thống hỗ trợ ra quyết định đối với ứng dụng trong nông nghiệp [117];
- Thực hiện kiểm tra ô tô không người lái [118], [119];
- Kiểm tra hệ thống nhận dạng lời nói và giọng nói [120], [121];
- Người máy trong lĩnh vực chăm sóc sức khỏe [122];
- Đo lường khối lượng công việc nhận thức của chatbot (trợ lý giọng nói v.v..) [123]; và
- Kiểm tra hệ thống dạy kèm thông minh [124].

Các thử nghiệm hiện trường cho các hệ thống AI khác nhau rất nhiều về phương pháp luận, số lượng người dùng, các trường hợp sử dụng liên quan, tình trạng của tổ chức/người chịu trách nhiệm và tài liệu hóa các kết quả. Cho dù các thử nghiệm có thể được áp dụng như một biện pháp để cải thiện chất lượng của các hệ thống AI phụ thuộc vào các rủi ro liên quan đến việc áp dụng các hệ thống đó. Trong nhiều ứng dụng, thử nghiệm A/B được sử dụng như một kỹ thuật để cung cấp các phiên bản hệ thống khác nhau cho những người dùng khác nhau nhằm so sánh hiệu năng của hệ thống.

Ngoài tính nghiêm ngặt một cách hợp lý của phần mềm AI được xây dựng, cần tính đến khả năng chấp nhận đối với con người và các thử nghiệm hiện trường có thể giúp đạt được điều đó. Thêm vào đó lỗi của hệ thống AI trong một bài kiểm tra chức năng có thể là không cần thiết hoặc có thể không thể giải quyết được. Hệ thống AI hiển thị các kết quả thay đổi có thể được coi là hữu ích cho mục đích dự kiến của chúng, khả năng của hệ thống AI đạt được kết quả theo kế hoạch và mong muốn không phải lúc nào cũng có thể đo lường được bằng các phương pháp kiểm thử phần mềm thông thường.

Một sự khác biệt cơ bản khác giữa nhiều hệ thống AI và các hệ thống thông thường là hệ thống thông thường được thiết kế để phát triển, sản xuất và kiểm soát chất lượng nhằm đáp ứng nghiêm ngặt các thông số kỹ thuật nhất định. Phần mềm truyền thống được thiết kế để có thể tái tạo các hành vi của nó, trong khi các hệ thống AI là sự tìm kiếm khả năng tổng quát hóa. Điều này dẫn đến việc thực nghiệm kiểm tra và thử nghiệm hiện trường có thể hiệu quả hơn để đánh giá chất lượng.

Làm thế nào để đối phó với sự không chắc chắn về kết quả của sản phẩm và rủi ro khi triển khai sản

phẩm là đối tượng đề cập của nhiều quy định trong lĩnh vực y tế. Hệ thống AI y tế có thể yêu cầu phải tuân thủ ISO 14155 [125]. Chúng có thể phải trải qua "điều tra lâm sàng", một quy trình tương tự như "thử nghiệm lâm sàng" [126] [127]. Điều này cũng đúng với các lĩnh vực khác, chẳng hạn như hệ thống cho hạt nhân hay điều khiển chuyến bay.

9.10.2.7 So sánh với trí thông minh của con người

Trong trường hợp hệ thống AI được thiết kế để tự động hóa hoạt động của con người liên quan đến việc xử lý dữ liệu và ra quyết định, một trong những cách để thẩm định hệ thống AI là so sánh với khả năng thông minh của con người. Cách tiếp cận như vậy có thể cho phép các bên liên quan khác nhau như người dùng chính của hệ thống AI, các đối tượng bên ngoài và cơ quan xây dựng chính sách trong thuộc lĩnh vực triển khai của AI (bao gồm cả các cơ quan quản lý) tin tưởng việc hệ thống AI thực hiện được một số tác vụ ứng dụng liên quan đến xử lý dữ liệu và ra quyết định, cái mà trước đây chủ yếu do con người thực hiện.

Ví dụ hữu ích về việc so sánh với khả năng của con người là các hoạt động được cấp phép theo truyền thống, chẳng hạn như điều khiển phương tiện cơ giới hoặc chăm sóc sức khỏe. Cho phép xe tự hành chạy trong thành phố đường phố hoặc một hệ thống tự quản để thực hiện bất kỳ biện pháp xử lý nào, sẽ chỉ được cho phép nếu có bằng chứng cho thấy hệ thống AI tiến hành các hoạt động này không tệ hơn con người. Cách tiếp cận như vậy cho phép những điều sau:

- Người dùng và các bên liên quan của hệ thống AI có thể mong đợi rằng chất lượng của hệ thống AI khi thực hiện tác vụ xử lý thông tin không kém hơn chất lượng đối với giải pháp cho cùng một vấn đề được thực hiện bởi người điều hành;
- Các bên thứ ba có thể mong đợi rằng hoạt động của hệ thống AI sẽ không gây nguy hiểm, thiệt hại về người và tài sản, vật chất.

Có thể kết luận rằng một hệ thống AI không kém hơn khả năng của con người khi thực hiện các tác vụ ứng dụng liên quan đến xử lý dữ liệu và đạt được sự an toàn trong xử lý, nếu các số liệu thống kê AI không kém hơn một giá trị ngưỡng đã xác định.

Để có được các giá trị ngưỡng và sự tin tưởng như vậy về an toàn trong xử lý của các hệ thống AI, điều quan trọng là sử dụng các mẫu dữ liệu đại diện phản ánh được bản chất của tác vụ xử lý thông tin để có thể áp dụng cho cả hệ thống AI hoặc trí tuệ tự nhiên của con người.

9.10.3 Các quan tâm về độ bền vững

Định nghĩa về độ bền vững là "khả năng của một hệ thống duy trì mức độ thực thi của nó trong bất kỳ hoàn cảnh nào". Để hiểu độ bền vững theo nghĩa tổng quát, điều quan trọng cần lưu ý là những gì mà các hệ thống AI thường được sử dụng, ví dụ như suy luận tri thức (cách tiếp cận biểu trưng) hoặc tổng quát hóa từ dữ liệu (cách tiếp cận biểu trưng phụ).

Nguyên tắc chính là một hệ thống AI được mong đợi có thể hoạt động trên dữ liệu chưa được biết trước và trong các bối cảnh có thể có những thay đổi đáng kể. Một hệ thống AI được kỳ vọng sẽ đối phó với

các điều kiện làm việc có thể thay đổi rất nhiều và độ mạnh mẽ của nó tương ứng với khả năng tiếp tục hoạt động theo thiết kế của nó. Tùy thuộc vào loại hình hệ thống AI, cần có các số liệu đo lường khác nhau để đánh giá độ bền vững của hệ thống.

Khi một hệ thống AI được sử dụng để thực hiện phép nội suy, độ bền vững của nó được coi là "khả năng có thể chấp nhận phạm vi nhất định về biên độ phản hồi trên bất kỳ đầu vào hợp lệ nào". Điều này có nghĩa là hệ thống AI được kỳ vọng sẽ không thể hiện hành vi không ổn định trong phép nội suy của nó.

Khi một hệ thống AI được sử dụng để thực hiện phân loại, độ bền vững của nó được xem là "khả năng gán phân loại nhất quán cho cả đầu vào đã biết và đầu vào biến đổi trong một phạm vi nhất định". Điều này có nghĩa là hệ thống AI có thể tiến hành phân loại đúng cách cho cả đầu vào đã biết và chưa biết, miễn là chúng (chưa biết) không quá khác biệt so với đầu vào đã biết.

Khi một hệ thống AI được sử dụng để thực hiện một tác vụ giải quyết, độ bền vững của hệ thống này được coi là "khả năng đưa ra một giải pháp vẫn hiệu quả khi có một sự thay đổi có thể chấp nhận được của vấn đề cần giải quyết ban đầu". Điều này có nghĩa là hệ thống AI có thể tạo ra các giải pháp chấp nhận được cho các vấn đề cần giải quyết khác nhau miễn là chúng không quá khác biệt so với vấn đề cần giải quyết ban đầu.

Khi một hệ thống AI được sử dụng để thực hiện tính điểm, độ bền vững của nó được coi là "khả năng chỉ định các phương pháp đo lường tin cậy và nhất quán về xếp hạng cho cả đầu vào đã biết và đầu vào thay đổi trong phạm vi chấp nhận được". Điều này có nghĩa là trong trường hợp đầu vào và đầu ra không xác định, hệ thống AI sẽ gán điểm về cơ bản sẽ không khác nhau đối với các đầu vào đã biết và đầu vào không xác định miễn là chúng không quá khác so với các đầu vào đã biết.

9.10.4 Các quan tâm liên quan đến quyền riêng tư

Để giải quyết các mối đe dọa về quyền riêng tư trong AI, các chỉ số về quyền riêng tư giúp đánh giá mức độ riêng tư và mức độ bảo vệ mà hệ thống cung cấp. Việc xác định và áp dụng các chỉ số về quyền riêng tư nhằm mục đích giải quyết thách thức này. Có nhiều kỹ thuật máy học bảo vệ quyền riêng tư hoặc tăng cường quyền riêng tư trong AI để bảo vệ dữ liệu nhạy cảm ở các miền khác nhau. Mục đích xác định các chỉ số về quyền riêng tư là để định lượng mức độ riêng tư của dữ liệu dẫn đến việc cải thiện mô hình quyền riêng tư trong một mô hình AI cụ thể [128][129]. Xét về mặt kỹ thuật, số liệu đo lường về quyền riêng tư xem xét các thuộc tính khác nhau của dữ liệu để từ đó có được giá trị đại diện cho mức độ riêng tư trong hệ thống. Ưu điểm của các số liệu đo lường về quyền riêng tư là nó cung cấp khả năng so sánh các kỹ thuật bảo vệ quyền riêng tư khác nhau, đánh giá các phương pháp bảo vệ quyền riêng tư khác nhau trong một miền cụ thể và để giảm thiểu bộc lộ các quyền riêng tư. Các chỉ số đo lường quyền riêng tư rất hữu ích khi dữ liệu nhạy cảm bị đe dọa bởi kẻ xấu. Các chỉ số về quyền riêng tư có sự khác biệt khi khi xét đến nguồn dữ liệu theo các khía cạnh đánh giá khác nhau về quyền riêng tư và các mối đe dọa.

9.10.5 Các quan tâm về khả năng dự đoán của hệ thống

Một số cách tiếp cận kiểm tra và thẩm định mô tả ở trên là cần thiết để đánh giá khả năng dự đoán của

hệ thống AI. Khả năng dự đoán có thể được đo lường thông qua phản hồi tường minh mang tính chủ quan chủ quan từ các cuộc điều tra, phỏng vấn dựa trên bảng câu hỏi, trong đó người tham gia được yêu cầu suy nghĩ để đưa ra các mục tiêu, dự đoán các hành động trong tương lai của người máy chẳng hạn [130]. Các chỉ số đo lường khác cũng có thể được sử dụng, chẳng hạn như thời gian phản ứng để người dùng nhận ra ý định của hệ thống AI và phản ứng tương ứng, chẳng hạn như thời gian phản ứng ngắn hơn cho thấy khả năng dự đoán cao hơn. Hành vi nhìn cũng cho thấy dấu hiệu gián tiếp về khả năng dự đoán của người máy, chẳng hạn như người tham gia phải nhìn người máy nhiều lần hoặc nhìn lâu hơn thì khả năng dự đoán của người máy càng kém [131]. Thử nghiệm hệ thống AI trên một số lượng lớn các tổ hợp điều kiện môi trường sẽ cho phép mô tả đầy đủ hơn về hành vi của nó. Dựa trên đặc điểm này người dùng biết những gì họ kỳ vọng ở hệ thống để tạo khả năng dự đoán của nó.

9.11 Sử dụng và khả năng áp dụng

9.11.1 Sự tuân thủ

Nhu cầu tuân thủ một phần do việc áp dụng các tiêu chuẩn và quy định khác nhau trong các ngành công nghiệp. Hệ thống AI cần phải tính đến tuân thủ các tiêu chuẩn và quy định hiện hành chứ không xem xét riêng rẽ trong từng trường hợp sử dụng.

9.11.2 Quản lý các kỳ vọng

Quản lý kỳ vọng là cần thiết để tránh phá vỡ lòng tin vì hệ thống không thể thực hiện các kỳ vọng không thực tế. Điều này đòi hỏi sự rõ ràng về khả năng thực tế của hệ thống AI, bao gồm phạm vi đầu vào để có được đầu ra chắc chắn, có thể tin cậy đúng như kỳ vọng (đặc biệt đối với các hệ thống dựa trên suy luận thống kê).

9.11.3 Ghi nhãn sản phẩm

Ghi nhãn sản phẩm và hệ thống (bao gồm các liên kết đến thông tin để cập nhật kịp thời) có thể cần thiết cho sự an toàn của cả người dùng và nhà cung cấp hệ thống AI:

- Rằng người dùng cuối đang tương tác với một tác nhân AI cũng như tuyên bố về ý định/mục đích của hệ thống AI;
- Các rủi ro và hạn chế của mô hình;
- Tần suất tái huấn luyện khi cần thiết;
- Ngày đánh giá hiệu năng cuối cùng được thực hiện;
- Nguồn và ngày sử dụng dữ liệu huấn luyện [132].

9.11.4 Nghiên cứu khoa học về nhận thức

Dựa trên thảo luận trong mục này, hướng dẫn cụ thể và mô tả các mối quan tâm là cần thiết phải có để đạt được mức độ đáng tin cậy hợp lý và sử dụng các hệ thống một cách đúng đắn. Tuy nhiên một số dạng được ghi nhận có sự tương tác giữa chất lượng của một hệ thống và khả năng bị lạm dụng hoặc không sử dụng. Ví dụ các hệ thống đáng tin cậy ở mức độ cao đã gây ra sự phụ thuộc quá mức dẫn đến

việc lạm dụng trong quá trình sử dụng (chẳng hạn như trong hàng không), trong khi các hệ thống không đáng tin cậy (như EHR) đã gây ra sự ngờ vực và do đó dẫn đến không sử dụng [133]. Một số dạng tương tự cũng áp dụng cho các chỉ số đo lường khác như độ bền, khả năng phục hồi và độ chính xác – hệ thống tốt hơn sẽ truyền cảm hứng cho sự tin tưởng, điều này có thể gây ra lỗi trong hệ thống ở các tình huống mà hệ thống tự động không được thiết kế để xử lý.

Do đó phải luôn nhìn nhận và duy trì các yếu tố liên quan đến con người, tốt nhất là nên có một cách nhìn đa diện về việc tối ưu hóa các chỉ số đo lường liên quan đến con người trong các hệ thống AI giao tiếp với con người. Quan điểm này sẽ dựa trên sự tương tác giữa con người với máy tính (HCI), nghiên cứu khoa học về nhận thức để thu được các kết quả hữu ích cũng như các tiêu chuẩn phù hợp về mặt khoa học.

10 Kết luận

Việc nhận ra những lợi ích tiềm năng của các hệ thống AI có thể bị cản trở bởi sự thiếu tin tưởng của khách hàng, người dùng và xã hội nói chung vào tính tin cậy, tính hiệu quả, công bằng và thậm chí là mục đích của các ứng dụng AI. Các mối quan tâm trong kinh doanh, chính phủ, xã hội và đạo đức nếu không được giải quyết một cách có hệ thống có thể làm xói mòn niềm tin vào AI. Những mối quan ngại như vậy có thể gia tăng vì tính dễ bị tổn thương thể hiện trong các hệ thống AI dựa trên ML, ví dụ như sự thiên vị, tính không thể đoán trước và sự không rõ ràng. Nhiều ứng dụng ML được thúc đẩy phát triển bởi nhu cầu khai phá dữ liệu lớn, quyền riêng tư dữ liệu và các vấn đề khác về quản trị dữ liệu, ví dụ như tính truy nguyên và chất lượng dữ liệu có thể trở thành mối quan tâm lớn trong việc xây dựng và sử dụng các hệ thống AI. Để cải thiện tính đáng tin cậy của AI thì tính dễ bị tổn thương của ML và dữ liệu cần được xem xét và giải quyết một cách rõ ràng trong các chính sách, quy trình và trên cơ sở từng trường hợp sử dụng.

Tác động tiềm ẩn của tính dễ bị tổn thương của AI đối với các bên liên quan cần được kiểm tra để quyết định xem liệu việc sử dụng AI có phù hợp trong các trường hợp cụ thể hay không. Một tổ chức đang phát triển hoặc sử dụng AI có thể áp dụng phương pháp tiếp cận dựa trên rủi ro để xác định các tác động có thể xảy ra đối với tổ chức, đối tác của tổ chức, đối với người dùng dự kiến và đối với xã hội để giảm thiểu rủi ro một cách phù hợp. Điều quan trọng là tất cả các bên liên quan phải hiểu bản chất của các rủi ro tiềm ẩn và các biện pháp giảm thiểu cần được thực hiện.

Các phép đo định lượng về chất lượng và các mô hình có thể lặp lại của các quá trình có thể góp phần tạo lập và duy trì niềm tin vào các hệ thống AI hiện vẫn chưa được xác lập. Điều cần thiết là phải áp dụng các phương pháp luận hiện có và mới nổi để cải thiện độ bền vững của các hệ thống AI để đạt được cấp độ nào đó theo chỉ định.

Tóm lại, tiêu chuẩn này trình bày về tính đáng tin cậy của các hệ thống và công nghệ AI dựa vào việc giải quyết những mối quan tâm của các bên liên quan về AI, sử dụng dữ liệu của nó một cách minh bạch và dễ tiếp cận, xây dựng các hệ thống AI có độ bền vững về mặt kỹ thuật, có thể kiểm soát và kiểm chứng được trong toàn bộ vòng đời của chúng.

Phụ lục A

(Tham khảo)

Nghiên cứu liên quan về các vấn đề xã hội

Đã có rất nhiều nghiên cứu ở đa lĩnh vực đề cập đến tính đáng tin cậy liên quan đến các vấn đề xã hội, chúng bao gồm các vấn đề về đạo đức trong công nghệ; đạo đức trong nghiên cứu và đổi mới (thường được gọi là nghiên cứu và đổi mới có trách nhiệm); các thuật toán có trách nhiệm giải trình; đạo đức trong sử dụng và xử lý dữ liệu; bảo vệ dữ liệu trong hoạt động quản trị dữ liệu. Trong nhiều trường hợp, mục đích và phạm vi của những nghiên cứu này ở cấp độ cao so hơn thay vì cung cấp các chuẩn cho phép các nhà phát triển, nhân viên, khách hàng, người dùng và xã hội nói chung xác định một cách hiệu quả và chính xác mức độ tin cậy mà họ sẽ đặt vào bất kỳ dịch vụ hoặc sản phẩm cụ thể nào sử dụng AI.

Ví dụ về nghiên cứu hiện có bao gồm:

- IEEE đã đưa ra những phân tích mang tính toàn diện xung quanh vấn đề thiết kế để phù hợp về mặt đạo đức đối với các hệ thống tự trị và thông minh [134]. Điều này nhằm mục đích phát triển đạo đức nghề nghiệp. Vì vậy nó cung cấp đầu vào phù hợp đối với các lĩnh vực yêu cầu trách nhiệm giải trình, tính minh bạch, khả năng xác minh, khả năng dự đoán, sự phù hợp với các chuẩn mực và giá trị xã hội cũng như các phương pháp thiết kế.
- Hiện có một nhóm xúc tiến mạnh mẽ các hoạt động nghiên cứu hướng đến sự hài hòa hóa trong nghiên cứu và đổi mới có trách nhiệm (RRI), trong đó đề cập đến rất nhiều nỗ lực nghiên cứu nhưng đều nhắm đến mục tiêu trước hết là sự hỗ trợ của quỹ tài trợ công [135]. Những nghiên cứu này đưa ra cách thức tiếp cận cơ bản để cung cấp các kỹ thuật cần thiết cho việc phát triển sự tin cậy vào AI, nhưng nó cũng bao gồm các vấn đề nằm ngoài phạm vi được đề cập chẳng hạn như công bố khoa học, đào tạo và cân bằng giới giữa các nhà nghiên cứu, dữ liệu nghiên cứu mở và thử nghiệm trên động vật. Kinh nghiệm liên quan đến việc áp dụng RRI trong phát triển do tư nhân tài trợ cũng sẽ là một ưu tiên [136][137].
- Mục tiêu của ISO/IEC AWI 38507 [138] là hỗ trợ cung cấp bối cảnh sử dụng cho việc quản trị các loại hình quyết định của AI.

Thư mục tài liệu tham khảo

[1] ISO/IEC 27000:2018, Information technology — Security techniques — Information security management systems — Overview and vocabulary

[2] ISO/IEC 11557:1992, Information technology — 3,81 mm wide magnetic tape cartridge for information interchange — Helical scan recording — DDS-DC format using 60 m and 90 m length tapes

[3] ISO/IEC 2382:2015, Information technology — Vocabulary

[4] ISO/IEC/IEEE 15939:2017, Systems and software engineering — Measurement process

[5] ISO/IEC 21827:2008, Information technology — Security techniques — Systems Security Engineering — Capability Maturity Model® (SSE-CMM®)

[6] IEC 61800-7-1:2015, Adjustable speed electrical power drive systems — Part 7-1: Generic interface and use of profiles for power drive systems — Interface definition

[7] ISO 5127:2017, Information and documentation — Foundation and vocabulary

[8] ISO 9000:2015, Quality management systems — Fundamentals and vocabulary

[9] ISO/IEC 10746-2:2009, Information technology — Open distributed processing — Reference model: Foundations — Part 2

[10] ISO/IEC Guide 51:2014, Safety aspects — Guidelines for their inclusion in standards

[11] ISO 13702:2015, Petroleum and natural gas industries — Control and mitigation of fires and explosions on offshore production installations — Requirements and guidelines

[12] ISO 10018:2020, Quality management — Guidance for people engagement

[13] ISO 18115-1:2013, Surface chemical analysis — Vocabulary — Part 1: General terms and terms used in spectroscopy

[14] ISO 31000, Risk management — Guidelines

[15] ISO 18646-2:2019, Robotics — Performance criteria and related test methods for service robots — Part 2: Navigation

[16] ISO 8373:2012, Robots and robotic devices — Vocabulary

[17] ISO/IEC 38500:2015, Information technology — Governance of IT for the organization

[18] ISO/IEC/IEEE 15288:2015, Systems and software engineering — System life cycle processes

[19] ISO/IEC 25012:2008, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model

[20] ISO/IEC 25010:2011, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models

[21] ISO/IEC/TR 22417:2017, Information technology — Internet of things (IoT) use cases

[22] ISO/IEC/IEEE 24765:2017, Systems and software engineering — Vocabulary

[23] ISO 22316:2017, Security and resilience — Organizational resilience — Principles and attributes

[24] ISO/IEC/TR 29110-1:2016, Systems and software engineering — Lifecycle profiles for Very Small Entities (VSEs) — Part 1: Overview

[25] ISO 10303-11:2004, Industrial automation systems and integration — Product data representation and exchange — Part 11: Description methods: The EXPRESS language reference manual

[26] United Nations Environment Programme (UNEP) Rio Declaration on Environment and Development, 1992 (http://www.jus.uio.no/lm/environmental.development.rio. declaration. 19 92/portrait.a4.pdf)

[27] ITU-T Y.3052, Overview of Trust Provisioning in ICT Infrastructures and Services, 2017 (https:// www.itu.int/rec/T-REC-Y.3052/en)

[28] IEEE 1012:2016, IEEE Standard for System, Software and Hardware Verification and Validation

[29] Mohammadi N. et al. Trustworthiness Attributes and Metrics for Engineering Trusted InternetBased Software Systems, 2014. Communications in Computer and Information Science. 453. 19- 35. 10.1007/978-3-319-11561-0_2 (https://www.researchgate. net/publication/267390448 _Trustworthiness _Attributes _and _Metrics _for _Engineering _Trusted _Internet -Based _Software _Systems)

[30] Colquitt J., Scott B., Le Pine J. A, Trust, Trustworthiness and Trust Propensity: A MetaAnalytic Test of Their Unique Relationships With Risk Taking and Job Performance, The Journal of applied psychology, 2007. 92. 909-27. 10.1037/0021-9010.92.4.909

(https://www.researchgate .net/publication/6200985 _Trust _Trustworthiness _and _Trust_ Propensity _A _Meta -Analytic_Test_of_Their_Unique_Relationships_With_Risk_Taking _and _Job_Performance)

[31] Marr B. What Is Deep Learning AI? A Simple Guide With 8 Practical Examples, 2018 (https://www .forbes .com/sites/bernardmarr/2018/10/01/what -is -deep -learning -ai -a -simple -guide -with -8 -practical-examples)

[32] O'Keefe K., Daragh O Brien, Ethical Data and Information Management: Concepts, Tools and Methods, Kogan Page pp. 46-47, 214-218, 262-263, 2018

[33] [33] Freeman R.E., McVea J. A Stakeholder Approach to Strategic Management, 2001, Darden Business School Working Paper No. 01-02

[34] The European Commission High Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AI, Working Document for stakeholder' consultation, Brussels, December 2018

[35] IEEE EAD v2, Ethically Aligned Design — Version II Request for Input, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2018

[36] [36] Borning A., Muller M. Next steps for Value Sensitive Design, 2012. Proceedings of CHI, 1125- 1134. New York, NY, ACM Press, 2012

[37] ISO/IEC 38505-1:2017, Information technology — Governance of IT — Governance of data — Part 1: Application of ISO/IEC 38500 to the governance of data

[38] Winikoff M. How to make robots that we can trust, 2018 (http://theconversation.com/how-to - make-robots-that-we-can-trust-79525)

[40] Doshi-Velez F. et al. Accountability of AI Under the Law: The Role of Explanation, SSRN Electronic Journal, 2017

[41] European Group on Ethics in Science and New Technologies Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, 2018 (doi:10.2777/531856)

[42] ISO/IEC 30141:2018, Internet of Things (IoT) — Reference Architecture

[43] Commission de Surveillance du Secteur Financier Artificial Intelligence: opportunities, risks and recommendations for the financial sector, 2018

[44] Pei K. et al. Towards Practical Verification of Machine Learning: The Case of Computer Vision Systems, 2017

[45] Szegedy C. et al. Intriguing properties of neural networks, 2014

[46] Goodfellow I. et al. Explaining and Harnessing Adversarial Examples, 2015

[47] Brendel W., Rauber J., Bethge M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models, 2018

[48] Akhtar N., Mian A., Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey, 2018

[49] Yuan X. et al. Adversarial Examples: Attacks and Defences for Deep Learning, July 2018

[50] Raghunathan A., Steinhardt J., Liang P. Certified Defenses against Adversarial Examples, 2018

[51] ISO/IEC 29100:2011, Information technology — Security techniques — Privacy framework

[52] ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and classification of techniques

[53] Truta T.M., Vinay B. Privacy Protection: p-Sensitive k-Anonymity Property, 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. 94-94

[54] Rocher L., Hendrickx M., de Montjoye Y., Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications 10, Article number: 3069, 2019

[55] Bergman M., Van Zandbeek M. Close Encounters of the Fifth Kind? Affective Impact of Speed and Distance of a Collaborative Industrial Robot on Humans, Human Friendly Robotics, p. 127-137. Springer, Cham, 2018

[56] Brownlee J. Ph.D., Discover Feature Engineering, How to Engineer Features and How to Get Good at It, Machine Learning Mastery, 2014 (https://machinelearningmastery.com/discover-feature -engineering-how-to-engineer-features-and-how-to-get-good-at-it/)

[57] Parasuraman R., Riley V. A., Humans and automation: Use, misuse, disuse, abuse, Human Factors, vol. 39, pp. 230–253

[58] Haines T. S.F., Mac Aodha O., Brostow G. J. My Text in Your Handwriting, University College London, Transactions on Graphics, 2016 (http://visual.cs.ucl.ac.uk/pubs/handwriting/)

[59] Van den Oord A. WaveNet: A Generative Model for Raw Audio, 2016 (https://deepmind.com/ blog/wavenet-generative-model-raw-audio/)

[60] Wagner K. Someone built chatbots that talk like the characters from HBO's 'Silicon Valley', 2016 (https://www.recode .net/2016/4/24/11586346/silicon -valley -hbo -chatbots -for -season -3 -premier)

[61] Dormon B. The AI that (almost) lets you speak to the dead, 2016 (https://arstechnica.com/ information-technology/2016/07/luka-ai-chatbot-speaking-to-the-dead-mind-uploading/).

[62] Newitz A. Ashley Madison code shows more women and more bots, 2015 (https://gizmodo.com/ ashley-madison-code-shows-more-women-and-more-bots-1727613924)

[63] World Economic Forum & UNICEF Innovation Centre Generation AI, (https://www.weforum.org/ projects/generation-ai, https://www.unicef.org/innovation/stories/ generation-ai)

[64] Pearl J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, ArXiv:1801.04016 [Cs, Stat], January 2018 (http://arxiv.org/abs/1801.04016)

[65] Guidotti R. et al., A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51:1–42, 2019

[66] [66] Achinstein P., The Nature of Explanation, 1985, Oxford University Press, 2017 (https:// books.google. fi/books?id=TkULPY-xMNsC)

[67] Adadi A., Berrada M. 2018, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access; 6:52138-60

[68] Doshi-Velez F., Kim B. 2017, Towards A Rigorous Science of Interpretable Machine Learning, arXiv 1702.08608

[69] Lipton Z. C., The mythos of model interpretability, Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018

[70] Dhurandhar A. et al. 2017, TIP: Typifying the Interpretability of Procedures, arXiv preprint arXiv:1706. 02952

[71] Miller T. Explanation in artificial intelligence: insights from the social sciences, 2017, arXiv preprint arXiv:1706. 07269

[72] https://pair-code.github. io/facets/

[73] https://projector.tensorflow.org/

[74] Holland S. et al. 2018, The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards, arXiv preprint arXiv: 1805. 03677

[75] Gebru T. et al. 2018, Datasheets for Datasets, arXiv preprint arXiv: 1803. 09010

[76] Seck I. et al. 2018, Baselines and a datasheet for the Cerema AWP dataset, arXiv preprint arXiv: 1806 .04016

[77] Angelino E. et al. 2018, Learning certifiably optimal rule lists for categorical data, Journal of Machine Learning Research, 18(234), pp.1-78

[78] Vaughan J. et al. 2018, Explainable Neural Networks based on Additive Index Models, arXiv preprint arXiv:1806. 01933

[79] Kim B., Khanna R., Koyejo O. 2016, Examples are not enough, learn to criticize! criticism for interpretability, Proc 30th Int Conf Neural Inf Process Syst: 2288–96

[80] Ribeiro M.T., Singh S., Guestrin C. 2016, August, Why should I trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144), ACM

[81] Kim B. et al. 2018, July, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), International Conference on Machine Learning (pp. 2673-2682)

[82] Koh P.W., Liang P. 2017, Understanding black-box predictions via influence functions, arXiv preprint arXiv:1703. 04730

[83] Maclaurin D., Duvenaud D., Adams R. 2015, June, Gradient-based hyperparameter optimization through reversible learning, International Conference on Machine Learning (pp.

2113-2122)

[84] Friedman J.H. 2001, Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232

[85] Lah C., Mordvintsev A., Schubert L. 2017, Feature visualization, Distill, 2(11), p.e7

[86] Olah C. et al. 2018, The building blocks of interpretability, Distill, 3(3), p.e10

[87] Erhan D. et al. 2009, Visualizing higher-layer features of a deep network, University of Montreal, 1341(3), p.1

[88] Lundberg S.M., Lee S.I. 2017, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems (pp. 4765-4774)

[89] Hohman F.M. et al. 2018, Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers, IEEE Transactions on Visualization and Computer Graphics

[90] Yosinski J. et al. 2015, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506. 06579

[91] Strobelt H. et al. 2018, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, IEEE transactions on visualization and computer graphics, 24(1), pp.667-676

[92] Strobelt H. et al. 2018, Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models, arXiv preprint arXiv: 1804. 09299

[93] Andrews R., Diederich J., Tickle A.B. 1995, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems, 8(6), pp.373-389

[94] Ailesilassie T. 2016, Rule extraction algorithm for deep neural networks: A review, arXiv preprint arXiv:1610. 05267

[95] Ribeiro M.T., Singh S., Guestrin C. 2018, Anchors: High-precision model-agnostic explanations, AAAI Conference on Artificial Intelligence

[96] List C. Levels: descriptive, explanatory and ontological, 2017 (http://philsci-archive.pitt.edu/ 13311/)

[97] Woodward J. In Making Things Happen: A Theory of Causal Explanation. Oxford Studies in the Philosophy of Science. Oxford, New York: Oxford University Press, 2005

[98] Montavon G., Samek W., Müller K.R. 2017, Methods for interpreting and understanding deep neural networks, Digital Signal Processing

[99] Lundberg S.M., Lee S.I. 2017, A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765-4774

[100] Gilpin L.H. et al. 2018, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning' arXiv preprint arXiv: 1806. 00069

[101] Narayanan M. et al. 2018, How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation, arXiv preprint arXiv: 1802.00682

[102] Hooker S. et al. 2018, Evaluating Feature Importance Estimates, arXiv preprint arXiv:1806.10758

[103] Samek W. et al. 2017, Evaluating the visualization of what a deep neural network has learned, IEEE transactions on neural networks and learning systems, 28(11), pp. 2660-2673

[104] Arras L. et al. 2017, What is relevant in a text document?: An interpretable machine learning approach, PloS one, 12(8), p.e0181142

[105] Bach S. et al. 2015, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one, 10(7), p.e0130140

[106] ISO/IEC 27040:2015, Information technology — Security techniques — Storage security

[107] IEC 61508:2010, Commented version, Functional safety of electrical/electronic/programmable electronic safety-related systems

[108] ISO 26262 (all parts), Road vehicles — Functional safety

[109] IEC 62279:2015, Railway applications — Communication, signalling and processing systems - Software for railway control and protection systems

[110] IEC 61511:2018, Functional safety — Safety instrumented systems for the process industry sector

[111] ISO/IEC/IEEE 29119-1:2013, Software and systems engineering — Software testing — Part 1: Concepts and definitions

[112] Zhou ZQ., Sun L. 2019, Metamorphic testing of driverless cars, Communications of the ACM 62:61–67

[113] Sojda R.S. 2007, Empirical evaluation of decision support systems: Needs, definitions, potential methods and an example pertaining to waterfowl management, Environmental Modelling & Software 22:269–277

[114] Ngan M., Grother P. 2015, Face Recognition Vendor Test (FRVT) - Performance of Automated Gender Classification Algorithms, National Institute of Standards and Technology

[115] Kohonen, Barna, Chrisley, 1988, Statistical pattern recognition with neural networks: benchmarking studies, IEEE 1988 International Conference on Neural Networks. pp 61–68

[116] BSI An investigation into the performance of facial recognition systems relative to their planned use in photo identification documents – BioP I. Bundesamt für Sicherheit in der Informationstechnik (BSI), Bundeskriminalamt (BKA), secunet Security Networks AG. 2004

(https:// www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Studies/BioP/BioPfinalreport _pdf)

[117] Burke J.J., Dunne B., Field testing of six decision support systems for scheduling fungicide applications to control Mycosphaerella graminicola on winter wheat crops in Ireland, The Journal of Agricultural Science, V 146 N 4, 2008

[118] UK-Government The Pathway to Driverless Cars: A Code of Practice for testing. Great Minster House, 33 Horseferry Road, London SW1 P 4DR: Department for Transport, 2015 (https:// assets.publishing.service.gov.uk/government/uploads/system/ uploads/attachment_data/file/ 446316/pathway-driverless-cars.pdf)

[119] Dressler F. et al. Inter-vehicle communication: Quo vadis, IEEE Communications Magazine, vol. 52, no. 6, pp. 170-177, June 2014 (doi: 10.1109/MCOM.2014.6829960)

[120] Lamel L. et al. Field trials of a telephone service for rail travel information. In Proceedings Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, 111–116. IEEE, 1996

[121] Isobe T. et al. Voice-activated home banking system and its field trial. In Proceedings Fourth International Conference on Spoken Language, ICSLP 96, 1688–1691. IEEE 1996.

[122] Jayawardena C. et al. Socially Assistive Robot HealthBot: Design, Implementation and Field Trials, in IEEE Systems Journal, vol. 10, no. 3, pp. 1056-1067, Sept. 2016 (doi: 10.1109/JSYST.2014.2337882)

[123] Strayer D. L. e t al. The smartphone and the driver's cognitive workload: A comparison of Apple, Google and Microsoft's intelligent personal assistants. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 71(2), 93, 2017

[124] Causo A. et al. Design of robots used as education companion and tutor. In Robotics and Mechatronics (pp. 75-84). Springer, Cham, 2016

[125] DIN EN ISO 14155:2012-01, Clinical investigation of medical devices for human subjects — Good clinical practice (ISO 14155:2011 + Cor. 1:2011); German version EN ISO 14155:2011 + AC: 2011, p.1–68, (https://www.beuth.de/de/norm/din-en-iso-14155/134954877)

[126] Läkemedelsverket The Medical Products Agency's Working Groupon Medical Information Systems. Medical Products Agency (Läkemedelsverket) Sweden, 2009 (https://lakemedelsverket .se/upload/foretag/medicinteknik/en/Medical-Information-Systems-Report_2009-06-18.pdf)

[127] Florek H.-J. et al. Results from a First-in-Human Trial of a Novel Vascular Sealant. Frontiers in Surgery, V 2, 2015

[128] Wagner I., Eckhoff D. Technical privacy metrics: a systematic survey, ACM Computing Surveys

(CSUR) 51.3 (2018): 57, 2018

[129] Wu F.-J. et al. The privacy exposure problem in mobile location-based services, Global Communications Conference (GLOBECOM), IEEE. IEEE, 2016

[130] Lichtenthäler C., Kirsch A. Goal-predictability vs. trajectory-predictability, which legibility factor counts. In proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, 2014. p. 228-229

[131] Lichtenthäler C., Lorenz T., Kirsch A. Towards a legibility metric: How to measure the perceived value of a robot. In International Conference on Social Robotics, ICSR 2011. 2011

[132] Davenport J.H. The debate about "algorithms", Mathematics Today 162, August 2017, pp. 162-165. (https://ima.org.uk/6910/the-debate-about-algorithms/)

[133] Wohleber R. et al. The Impact of Automation Reliability and Operator Fatigue on Performance and Reliance, 2016

[134] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016 (http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html)

[135] Reijers W. et al. Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations, Science and Engineering, Ethics Journal, pp 1-45, Springer, 2017

[136] Gurzawska A., Mäkinen M., Brey P., Implementation of Responsible Research and Innovation (RRI) Practices in Industry: Providing the Right Incentives, Sustainability 2017, 9, 1759 (doi:10.3390/su9101759)

[137] Lubberink R. et al., Lessons for Responsible Innovation in the Business Context: A Systematic Literature Review of Responsible, Social and Sustainable Innovation Practices, Sustainability, 2017, 9, 721 (doi:10.3390/su9050721)

[138] ISO/IEC AWI 38507, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.
