TCVN

TIÊU CHUẨN QUỐC GIA

TCVN DIN SPEC 92001-2:202x

TRÍ TUỆ NHÂN TẠO – QUY TRÌNH VÒNG ĐỜI VÀ YÊU CẦU CHẤT LƯỢNG – PHẦN 2: ĐỘ BỀN VỮNG

Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 2: Robustness

(Tài liệu Nghiệm thu Bộ)

HÀ NỘI - 2022

Lời nói đầu

TCVN DIN SPEC 92001-2:202x được xây dựng trên cơ sở tham khảo tài liệu DIN SPEC 92001-2 (2020) “Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness” của DIN SPEC (PAS).

TCVN DIN SPEC 92001-2:202x do Viện Khoa học Kỹ thuật Bưu điện - Học viện Công nghệ Bưu chính Viễn thông 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.

MỤC LỤC

MỤC LỤC **1**

1 Phạm vi áp dụng **2**
1.1 Lĩnh vực áp dụng 2
1.2 Giới hạn 2
1.3 Giới thiệu 2

2 Tài liệu viện dẫn **3**

3 Thuật ngữ và định nghĩa **3**
3.1 Thuật ngữ chung (General Terminology) 3
3.2 Thuật ngữ - Độ bền vững trước đối thủ (Terminology – Adversarial Robustness) 3
3.3 Thuật ngữ - Sai lệch độ bền vững (Terminology – Corruption Robustness) 5

4 Siêu mô hình chất lượng AI **8**

5 Độ bền vững **9**
5.1 Giới thiệu độ bền vững AI 9
5.2 Yêu cầu và hướng dẫn về quản lý rủi ro 10
5.3 Yêu cầu cụ thể với độ bền vững trước đối thủ 15
5.4 Yêu cầu cụ thể với sai lệch độ bền vững 27

6 Hướng dẫn thực hiện **36**

Tham chiếu đến các tài liệu về độ bền vững của AI tập trung được đề cập trong Thông số kỹ thuật DIN này: **38**

THƯ MỤC TÀI LIỆU THAM KHẢO **39**

Trí tuệ nhân tạo – Quy trình vòng đời và yêu cầu chất lượng – Phần 2: Độ bền vững

Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 2: Robustness

1 Phạm vi áp dụng

1.1 Lĩnh vực áp dụng

DIN SPEC này áp dụng cho tất cả các giai đoạn vòng đời của mô-đun AI - khái niệm, phát triển, triển khai, vận hành và ngừng hoạt động - và giải quyết nhiều quy trình vòng đời khác nhau. Do trong thực tế, các công nghệ AI được sử dụng cho nhiều nhiệm vụ khác nhau, nên DIN SPEC này không chỉ nhắm đến một lĩnh vực cụ thể mà còn áp dụng cho các công ty và sản phẩm AI trên tất cả các lĩnh vực.

DIN SPEC này áp dụng cho tất cả các loại mô-đun AI bao gồm ML và các hệ chuyên gia.

1.2 Giới hạn

Tiêu chuẩn này không định nghĩa hoặc liệt kê các thuật toán, phương pháp hoặc công nghệ là một phần của AI. Do đó, người dùng DIN SPEC được yêu cầu đánh giá xem siêu mô hình chất lượng AI quy định và các yêu cầu chất lượng AI liên quan có được áp dụng hay không.

Mặc dù yêu cầu đánh giá hành vi đạo đức trong phát triển mô-đun AI, DIN SPEC này không cung cấp bất kỳ yêu cầu cụ thể nào ở đó xác định hành vi đạo đức.

Các cân nhắc vòng đời phần mềm trong tiêu chuẩn này tương thích với ISO/IEC/IEEE 12207:2017, Hệ thống và công nghệ phần mềm - Các quy trình vòng đời phần mềm [1].

DIN SPEC đề xuất tách biệt giữa các mô-đun AI có rủi ro cao và thấp liên quan đến an toàn, bảo mật, quyền riêng tư và đạo đức. Nó cũng cung cấp các khía cạnh liên quan trong bối cảnh đánh giá rủi ro. Tiêu chuẩn này không thiết lập một quy trình đánh giá rủi ro chính xác, mà cũng không thiết lập khuôn khổ thiết kế đạo đức. Tuy nhiên, nó bị hạn chế bởi khung pháp lý hiện hành và quy tắc ứng xử đạo đức của mỗi tổ chức. Sự tuân thủ các quy định được cho là đúng. Ngoài ra, các bên liên quan của tiêu chuẩn này được yêu cầu thành lập một nhóm chuyên gia để đánh giá hồ sơ rủi ro mô-đun AI của họ.

Các yêu cầu chất lượng được liệt kê trong DIN SPEC 92001-2 không dành riêng cho từng miền. Các yêu cầu đã đưa ra trong tiêu chuẩn này được mở rộng cho các lĩnh vực ứng dụng AI cụ thể trong bước tiêu chuẩn hóa tiếp theo.

1.3 Giới thiệu

Trí tuệ nhân tạo (AI) là một lĩnh vực liên ngành và phức tạp. Nó cung cấp một công cụ để giải quyết các nhiệm vụ thường liên quan đến trí thông minh của con người. Mặc dù thực tế là các nguyên tắc vận hành chi phối một số phương pháp AI vẫn là một lĩnh vực nghiên cứu tích cực, nhưng hiệu suất của các mô-đun AI thường vượt trội so với phần mềm truyền thống đã dẫn đến sự gia tăng triển khai các công nghệ dựa trên AI.

Tuy nhiên, việc chuyển từ lĩnh vực nghiên cứu sang công nghệ có giá trị kinh tế và xã hội đòi hỏi phải thiết lập một khái niệm về chất lượng và độ tin cậy. Trong AI, không chỉ các vấn đề về chất lượng phần mềm truyền thống cần được xem xét mà cả những điều mới lạ như thiếu hiểu biết sâu sắc về logic cơ bản của mô-đun AI vốn là đặc trưng đối với Máy học (ML) trường phụ cũng cần được giải quyết. Hơn nữa, một số mô-đun AI thay đổi trong quá trình hoạt động

và do đó phải được theo dõi liên tục để việc xác thực và kiểm tra tính hợp lệ của AI không thể được coi là hoàn thành sau khi phát triển.

Do đó, tiêu chuẩn này xác định các yêu cầu vượt xa các yêu cầu chất lượng phần mềm truyền thống như được định nghĩa trong [1]. DIN SPEC 92001-1 thiết lập siêu mô hình chất lượng AI và vòng đời cho các mô-đun AI để làm nổi bật các đặc tính chất lượng của các mô-đun AI. Trong DIN SPEC 92001-2, tiêu chuẩn này, các yêu cầu cụ thể đảm bảo chất lượng AI liên quan đến độ bền vững được cung cấp. Cụ thể, trụ cột chất lượng AI "độ bền vững" được giải thích thêm và các yêu cầu cụ thể đối với trụ cột này được liệt kê. Hơn nữa, mỗi yêu cầu trong các trụ cột được ánh xạ tới một tập hợp các giai đoạn vòng đời để tạo thuận lợi cho việc phân loại các yêu cầu theo thời gian. Các yêu cầu về độ bền vững trong tiêu chuẩn này đã được phát triển với mục đích toàn diện về cả độ bền vững trước đối thủ và sai lệch độ bền vững nhằm hỗ trợ phát triển và triển khai các đổi mới AI an toàn và bảo mật.

2 Tài liệu viện dẫn

Không có tài liệu tham khảo viện dẫn trong tiêu chuẩn này.

3 Thuật ngữ và định nghĩa

Đối với các mục đích của tiêu chuẩn này, các thuật ngữ và định nghĩa sau đây được áp dụng. ISO, IEC và IEEE duy trì cơ sở dữ liệu thuật ngữ để sử dụng trong tiêu chuẩn hóa tại các địa chỉ sau:

- IEC Electropedia: có sẵn tại http://www.electropedia.org/.

- Nền tảng trình duyệt trực tuyến ISO: có sẵn tại http://www.iso.org/obp.

3.1 **Thuật ngữ chung** (General Terminology)

Tiêu chuẩn này dựa trên các thuật ngữ và định nghĩa của [1].

3.1.1

Trí tuệ nhân tạo (AI) (artificial intelligence - AI)

Lĩnh vực gồm nhiều ngành học thuật, thường được coi là một nhánh của khoa học máy tính, xử lý các mô hình và hệ thống để thực hiện các chức năng chung liên quan đến trí thông minh của con người, chẳng hạn như lý luận và kiến thức [2].

3.1.2

Mô-đun trí tuệ nhân tạo (AI) (artificial intelligence (AI) module)

Mô-đun phần mềm bao gồm các thuật toán AI.

3.2 **Thuật ngữ - Độ bền vững trước đối thủ** (Terminology – Adversarial Robustness)

3.2.1

Tấn công của đối thủ (adversarial attack)

Một thuật toán được phát triển bởi một đối thủ theo đó trả về các nhiễu loạn của đối thủ hoặc các mẫu của đối thủ.

3.2.2

Khả năng của đối thủ (adversarial capabilities)

Các hành động, thông tin, kỹ thuật hoặc các hướng tấn công khác nhau có thể dùng được để tấn công về mặt đe dọa [3].

3.2.3

Mẫu của đối thủ (adversarial example)

Đầu vào mô-đun AI mà kẻ tấn công đã cố tình thiết kế để gây nên mô hình mắc lỗi [4].

3.2.4

Khả năng chuyển giao mẫu của đối thủ (adversarial example transferability)

Thuộc tính mà các mẫu của đối thủ tạo ra để gây ra hoạt động sai trong một mô hình tương tự như gây ra hoạt động sai trong một mô hình khác [3].

3.2.5

Nhiễu loạn của đối thủ (adversarial perturbation)

Nhiễu loạn được tạo ra cẩn thận ở đó được thiết kế để gây ra hoạt động sai khi được thêm vào hoặc kết hợp với một hoặc nhiều điểm dữ liệu đầu vào của mô hình AI.

3.2.6

Độ bền vững trước đối thủ (adversarial robustness)

Khả năng của một mô-đun AI đối phó với các mẫu của đối thủ (hoặc nhiễu loạn của đối thủ).

3.2.7

Đối thủ (adversary)

Một cá nhân độc hại, người muốn tấn công mô-đun AI hoặc trong khi học bằng cách can thiệp vào dữ liệu huấn luyện, hoặc trong quá trình suy luận bằng cách thao túng các đầu vào trên mô hình đang đưa ra dự đoán [5].

3.2.8

Kiến thức của đối thủ (adversary's knowledge)

Kiến thức của đối thủ về mô-đun AI, ví dụ: bao gồm dữ liệu huấn luyện, kiến trúc mô hình, siêu tham số, số lớp, hàm kích hoạt, trọng số mô hình [6].

3.2.9

Phát hiện tấn công (attack detection)

Hành động phân biệt giữa hành vi bất thường và bình thường, hoặc giữa một mẫu của đối thủ và một mẫu lành tính [3].

3.2.10

Vi phạm tính khả dụng (availability violation)

Sự thỏa hiệp của các chức năng hệ thống bình thường có sẵn cho người dùng hợp pháp, chẳng hạn như độ chính xác, chất lượng hoặc quyền truy cập, dẫn đến đầu ra mô hình không thể truy cập hoặc không thể sử dụng được [3].

3.2.11

Mô hình hộp đen (blackbox model)

Đối thủ có kiến thức hạn chế hoặc không có kiến thức về mô hình bị tấn công và có thể không được phép thăm dò hoặc truy vấn mô hình trong khi xây dựng các mẫu đối thủ [3].

3.2.12

Chiến lược phòng thủ (defence strategy)

Một tập hợp các kỹ thuật giảm thiểu hoặc biện pháp đối phó để bảo vệ mô-đun AI chống lại

các tấn công của đối thủ.

3.2.13

Phạm vi nhiễu loạn cá nhân (individual perturbation scope)

Thuộc tính nhiễu loạn của đối thủ, cụ thể là nhiễu loạn được liên kết với một điểm đầu vào cụ thể và cố gắng đánh lừa mô-đun AI trên điểm đầu vào đã chọn này.

3.2.14

Vi phạm tính toàn vẹn (integrity violation)

Để tạo ra một đầu ra hoặc hành vi cụ thể theo lựa chọn của đối thủ [3].

3.2.15

Tấn công phi mục tiêu (non-targeted attack)

Một cuộc tấn công gây ra bất kỳ sự phân loại sai nào trái ngược với gây ra sự phân loại vào một lớp cụ thể (không chính xác) [3].

3.2.16

Vi phạm quyền riêng tư (privacy violation)

Tiết lộ thông tin cá nhân về sự riêng biệt có trong dữ liệu huấn luyện [3].

3.2.17

Tấn công có chủ đích (targeted attack)

Đối thủ cố gắng tạo ra các đầu vào theo đó buộc đầu ra của mô hình phân loại phải là một lớp mục tiêu cụ thể [7].

3.2.18

Phạm vi nhiễu loạn tổng thể (universal perturbation scope)

Thuộc tính nhiễu loạn của đối thủ, cụ thể là nhiễu loạn không thể biết trước hình ảnh và cố gắng đánh lừa mô-đun AI trên phần lớn các đầu vào [7].

3.2.19

Mô hình hộp trắng (whitebox model)

Đối thủ có kiến thức đầy đủ về mô hình bao gồm loại mô hình, kiến trúc mô hình và giá trị của tất cả các tham số và trọng số có thể huấn luyện [8].

3.3 Thuật ngữ - Sai lệch độ bền vững (Terminology – Corruption Robustness)

3.3.1 Sai lệch độ bền vững – Thuật ngữ chung (Corruption Robustness - General terminology)

3.3.1.1

Bất thường (anomaly)

Một mẫu trong dữ liệu không phù hợp với hành vi mong muốn [9].

3.3.1.2

Phát hiện bất thường (anomaly detection)

Các chiến lược và phương pháp khám phá sự bất thường và ngoại lệ [9].

3.3.1.3

Phát hiện thay đổi (change detection)

Các chiến lược và phương pháp khám phá những thay đổi trong các đầu vào đơn lẻ hoặc đầu vào phân tán [9].

3.3.1.4

Sai lệch dữ liệu (data corruption)

Quá trình dữ liệu được thay đổi từ định dạng ban đầu sang định dạng được coi là tạp nhiễu, không hoàn thiện hoặc khiếm khuyết [10].

3.3.1.5

Tập lỗi (error set)

Tập các điểm trong không gian đầu vào mà phân loại đưa ra dự đoán không chính xác [13].

3.3.1.6

Môi trường tạp nhiễu (noisy environment)

Môi trường của mô-đun AI nơi các biến được quan sát có một lỗi [14].

3.3.1.7

Môi trường không cố định (non-stationary environment)

Môi trường của mô-đun AI nơi phân phối dữ liệu cơ bản thay đổi theo thời gian [14].

3.3.1.8

Các ngoại lệ (outliers)

Một điểm dữ liệu khác biệt đáng kể so với các quan sát khác [15].

3.3.1.9

Tính nhất quán dự đoán (prediction consistency)

Một thuộc tính đảm bảo rằng, khi kích thước mẫu tăng lên, phân phối lấy mẫu của dự đoán trở nên ngày càng tập trung vào giá trị thực [10].

3.3.1.10

Độ tin cậy dự đoán (prediction reliability)

Một mô-đun AI được cho là có độ tin cậy dự đoán cao nếu nó đưa ra kết quả tương tự trong các điều kiện nhất quán [10].

3.3.1.11

Độ bền vững không gian (spatial robustness)

Độ bền vững của mô-đun AI đối với các phép biến đổi hình học, đặc biệt là phép tịnh tiến và phép quay [16].

3.3.1.12

Ẩn số chưa biết (unknown unknowns)

Các điểm dữ liệu không xảy ra trong dữ liệu huấn luyện hoặc kiểm tra của mô-đun AI (ở đó mô-đun AI không thể xử lý chính xác) [17].

3.3.2 Thay đổi phân phối/Thay đổi tập dữ liệu (Distributional Shift/ Dataset Shift)

3.3.2.1

Thay đổi phân phối/Thay đổi tập dữ liệu (distributional shift/dataset shift)

Phân phối ghép nối các đầu vào và các đầu ra khác nhau giữa giai đoạn huấn luyện hoặc kiểm tra và triển khai [11].

3.3.2.2

Thay đổi khái niệm (concept shift)

Sự thay đổi khái niệm xảy ra khi mối quan hệ giữa đầu vào và đầu ra thay đổi [12].

3.3.2.3

Thay đổi đồng biến (covariate shift)

Sự thay đổi phân phối của các biến đầu vào giữa giai đoạn huấn luyện hoặc kiểm tra và triển khai [12].

3.3.2.4

Thay đổi xác suất ưu tiên (prior probability shift)

Sự thay đổi phân phối biến đầu ra giữa giai đoạn huấn luyện hoặc kiểm tra và triển khai [12].

3.3.3 Sai lệch lựa chọn mẫu (Sample Selection Bias)

3.3.3.1

Sai lệch lựa chọn mẫu (sample selection bias)

Sự khác biệt giữa phân phối của dữ liệu huấn luyện hoặc thử nghiệmo và phân phối triển khai theo kết quả của quy trình từ chối mẫu không xác định [11].

3.3.3.2

Khiếm khuyết ngẫu nhiên (missing at random)

Khiếm khuyết ngẫu nhiên xảy ra khi phương pháp lấy mẫu phụ thuộc vào các đặc trưng đầu vào với điều kiện đã cho ở đó các đặc trưng đầu vào độc lập với đầu ra. Loại sai lệch này có tiềm năng tạo ra sự thay đổi đồng biến [12].

3.3.3.3

Khiếm khuyết ở lớp ngẫu nhiên (missing at random-class)

Khiếm khuyết ở lớp ngẫu nhiên xảy ra khi phương pháp lấy mẫu phụ thuộc vào đầu ra với điều kiện đã cho ở đó đầu ra độc lập với đầu ra của các đặc trưng đầu vào. Loại sai lệch này có tiềm năng tạo ra sự thay đổi xác suất ưu tiên [12].

3.3.3.4

Khiếm khuyết hoàn toàn ngẫu nhiên (missing completely at random)

Khiếm khuyết hoàn toàn ngẫu nhiên xảy ra khi phương pháp lấy mẫu hoàn toàn độc lập với các đặc trưng đầu vào và đầu ra [12].

3.3.3.5

Khiếm khuyết không ngẫu nhiên (missing not at random)

Khiếm khuyết không ngẫu nhiên xảy ra khi không có giả định độc lập giữa phương pháp lấy mẫu, các đặc trưng đầu vào và đầu ra. Loại sai lệch này có thể dẫn đến một hoặc nhiều thay đổi đồng biến, thay đổi xác suất ưu tiên và thay đổi khái niệm [12].

4 Siêu mô hình chất lượng AI

Siêu mô hình chất lượng AI được giới thiệu trong DIN SPEC 92001-1, Hình 1. Nó bao gồm các khía cạnh quan trọng nhất ở đó cần được tính đến để tạo thuận lợi cho thiết kế các mô-đun AI chất lượng cao. Vòng đời của một mô-đun AI, được minh họa bằng vòng tròn ở giữa Hình 1, bao gồm các giai đoạn vòng đời "khái niệm", "phát triển", "triển khai", "vận hành" và "ngừng hoạt động". Siêu mô hình chất lượng này thừa nhận ba nhóm quy trình vòng đời chính cần được xem xét trong các giai đoạn vòng đời. Điều này kích hoạt đề án có tổ chức, quản lý kỹ thuật và quy trình kỹ thuật. Các quy trình được gán cho ba nhóm này có thể được tìm thấy trong [14]. Các yêu cầu chất lượng đối với các mô-đun AI, thậm chí còn cao hơn so với phần mềm cổ điển, cần phải được liên kết với vòng đời. Điều này trở nên đặc biệt rõ ràng khi xem xét các ứng dụng ML phát triển thông qua thu thập thông tin khi triển khai trong thế giới thực, được biết đến như là học trực tuyến. Do đó, liên quan đến các yêu cầu nhất định, chẳng hạn như ngăn ngừa sai lệch có hại, được yêu cầu để tạo nguyên mẫu cũng như triển khai và vận hành vòng đời của mô-đun AI.

Hình 1 - Siêu mô hình chất lượng AI

Các đặc điểm chất lượng chính, được gọi là trụ cột chất lượng, cần được tính đến trong toàn bộ vòng đời của mô-đun AI là chức năng & hiệu năng, độ bền vững và tính dễ hiểu. Chúng được mô tả trong phần bên trái của Hình 1. Những trụ cột chất lượng này kéo theo những thách thức cấp bách nhất về chất lượng dành riêng cho AI. Chúng chỉ ra các vấn đề triển khai trung tâm của các mô-đun AI so với phần mềm cổ điển, chẳng hạn như xây dựng tập huấn luyện và độ bền vững tương ứng với các mẫu của đối thủ.

Trong DIN SPEC này, mỗi mô-đun AI được xem như hoặc là rủi ro cao hoặc là rủi ro thấp. Các mô-đun AI có liên quan đến an toàn, bảo mật, quyền riêng tư và đạo đức được phân loại như là các mô-đun có rủi ro cao (có tiềm năng). Các mô-đun AI không có sự liên quan như vậy được xem như rủi ro thấp. Đối với các mô-đun AI rủi ro cao, tiêu chuẩn này yêu cầu xem xét tất cả các yêu cầu chất lượng đã liệt kê. Những sai lệch tiềm tàng của các mô-đun AI rủi ro cao so với các yêu cầu chất lượng đã liệt kê cần có sự biện minh hợp lý. Yêu cầu này được nới lỏng đối với các mô-đun AI rủi ro thấp. Phân loại rủi ro này được thể hiện trong Bảng 1.

Trong Hình 1, đánh giá rủi ro cơ bản này được quan sát dưới dạng một hình chữ nhật được đặt phía sau ba trụ cột chất lượng. Bước đầu tiên trong giai đoạn khái niệm của mô-đun AI là đánh giá xem mô-đun AI được hỏi có tiềm ẩn rủi ro cao hay thấp liên quan đến các khía cạnh an toàn, bảo mật, quyền riêng tư và đạo đức hay không. Điều này không được xử lý trong DIN SPEC này. Một nhóm các chuyên gia trong một tổ chức chịu trách nhiệm về mô-đun AI sẽ được triệu tập cho nhiệm vụ này. Việc đánh giá rủi ro có ảnh hưởng đến tính nghiêm ngặt ở đó cần được áp dụng khi triển khai và đánh giá các thuộc tính và chất lượng của mô-đun AI.

Bảng 1 - Phân loại các yêu cầu liên quan đến các mô-đun AI rủi ro cao hoặc thấp theo bắt buộc, khuyến nghị cao và khuyến nghị

Lớp Mô-đun AI / Lớp yêu cầu	Rủi ro cao	Rủi ro thấp
Bắt buộc	Không sai lệch so với yêu cầu cho phép	Không sai lệch so với yêu cầu cho phép
Khuyến nghị mức cao	Sai lệch so với yêu cầu chỉ căn chỉnh	Sai lệch so với yêu cầu chỉ căn chỉnh
Khuyến nghị	Sai lệch so với yêu cầu chỉ căn chỉnh	Sai lệch so với yêu cầu chỉ căn chỉnh

5 Độ bền vững

5.1 Giới thiệu độ bền vững AI

Sau đây, độ bền vững sẽ biểu thị khả năng của mô-đun AI đối phó với dữ liệu đầu vào bị sai, tạp nhiễu, không xác định hoặc được xây dựng của đối thủ. Theo thực tế, mô-đun AI được đặt trong một môi trường phức tạp cao tiềm tàng (môi trường có thể không cố định về bản chất nhưng cũng có chiều cao về mặt quy trình tạo dữ liệu), độ bền vững là một vấn đề chất lượng AI quan trọng. Do đó, độ bền vững được giới thiệu và xử lý trong tiêu chuẩn này như một trụ cột cơ bản của chất lượng AI.

Trong tiêu chuẩn này, có sự phân biệt giữa độ bền vững trước đối thủ (AR), tức là độ bền vững đối với các nhiễu loạn của đối thủ và sai lệch độ bền vững (CR), tức là độ bền vững đối với các tín hiệu tạp nhiễu hoặc những thay đổi trong phân phối dữ liệu cơ bản. Hai mục tiêu độ bền vững cụ thể này cung cấp nội dung cho định nghĩa khá chung hơn về độ bền vững bằng cách bao hàm hai chủ đề và bộ tiêu chí thiết yếu về chất lượng AI.

AR yêu cầu các biện pháp bảo vệ mô-đun AI khỏi cái gọi là đối thủ đang cố đánh lừa mô-đun AI theo nghĩa đầu vào có hại được lựa chọn cẩn thận. Những năm nghiên cứu và phát triển AI gần đây nhất đã chỉ ra rằng các mô-đun AI tiên tiến nhất rất dễ bị tổn thương trước các mẫu bất lợi. Điều này đặt ra rủi ro nghiêm trọng khi các giải pháp AI được áp dụng trong các tình huống quan trọng về an toàn và bảo mật.

Mặt khác, CR yêu cầu các biện pháp làm giảm hoặc thậm chí ngăn ngừa các biến chứng phát sinh từ sự khác biệt giữa các tập dữ liệu được sử dụng trong giai đoạn phát triển và triển khai tương ứng. Có nhiều loại khác nhau thay đổi phân phối hoặc tập dữ liệu phải được giải quyết trong vòng đời AI, chẳng hạn như thay đổi đồng biến và thay đổi xác suất ưu tiên. Nguyên nhân nổi bật nhất của sự thay đổi phân phối là sai lệch lựa chọn và tính không ổn định của môi trường hoặc nền tảng.

Lưu ý rằng có một số biện pháp cải thiện cả AR và CR. Điều này là do thực tế cả AR và CR đều đảm bảo hiệu năng của mô-đun AI khi nó bộc lộ với các điểm dữ liệu được coi là rất khó xảy ra hoặc thậm chí là các điểm dữ liệu bên ngoài phân phối được nhìn thấy trong quá trình huấn luyện. Một mẫu của đối thủ thậm chí có thể được hiểu là một loại chuyển đổi phân phối đặc biệt ở đó được thiết kế đặc biệt để gây ra suy giảm hiệu suất của mô-đun AI.

Tuy nhiên, sự khác biệt vẫn có giữa AR và CR. Có nhiều lý do khác nhau cho điều này. Đáng chú ý nhất, vì một đối thủ hoạt động được giả định trong cài đặt AR, nên cần phải tính đến "cuộc chạy đua vũ trang" liên tục giữa tấn công và phòng thủ, trong khi CR thường được giải

quyết như một vấn đề phát sinh từ các nguyên nhân tự nhiên như hư hỏng phần cứng hoặc đầu vào bị xâm phạm. Do đó, AR và CR tương ứng được xem như vấn đề bảo mật và an toàn.

Trong khi CR tương ứng với và bị giới hạn trong các tình huống nơi không có thiết kế có chủ ý đằng sau các môi trường quan trọng dẫn đến suy giảm hiệu năng của mô-đun AI (do đó CR được gọi là "không tối ưu hóa"), AR là một lĩnh vực nghiên cứu hướng đến tối ưu hóa về bản chất. Tuy nhiên, cả CR và AR có thể phải tính đến, và được thử nghiệm trong ba môi trường thử nghiệm khác nhau, tức là miền kỹ thuật số, mô phỏng và vật lý.

Phần còn lại của tiêu chuẩn này được tổ chức như sau. Ở mức cao nhất, các điều 5.3 và 5.4 bao hàm những đóng góp chính của tiêu chuẩn này, tức là các yêu cầu liên quan đến AR và các yêu cầu liên quan đến CR. Mỗi điều được chia thành bốn khoản con được gọi là phân tích mô hình mối đe dọa, phân tích khả năng xảy ra và tác động, đánh giá độ bền vững và giảm nhẹ tổn hại. Mặc dù phân tích mô hình mối đe dọa hỗ trợ nâng cao hiểu biết hoặc nâng cao nhận thức về các lỗ hổng tiềm ẩn của một mô-đun AI, nhưng phân tích khả năng xảy ra và tác động sẽ định lượng khả năng xảy ra lỗi hiệu năng và thiệt hại hoặc chi phí liên quan. Đánh giá độ bền vững tiếp theo quy định một kế hoạch thử nghiệm có thể được tính đến như là công thức cuối cùng giảm nhẹ tổn hại.

Các yêu cầu truyền đạt các chi tiết quan trọng liên quan đến trụ cột chất lượng độ bền vững. Tiêu chuẩn này không hình thành một chuỗi hành động kiểm soát khi nào và cách thức các hành động được thực thi nhưng xa hơn bao gồm một tập hợp các yêu cầu chất lượng quan trọng về độ bền vững dành riêng cho AI. Tất cả các biện pháp kiểm soát chất lượng AI phải được nhúng vào các quy trình phù hợp để quản lý rủi ro cho một tổ chức đang phát triển hoặc sử dụng AI. Một tính toán đầy đủ về quản lý rủi ro nằm ngoài phạm vi của tiêu chuẩn này, nhưng mối liên kết chặt chẽ với các hoạt động quản lý rủi ro được công nhận và mô tả trong các yêu cầu và hướng dẫn về quản lý rủi ro tại điều 5.2. Khoản 5.2.2 đặt nền tảng phương pháp luận cho quản lý rủi ro tập trung vào AI bằng cách tập trung vào các nhu cầu an toàn và bảo mật cụ thể của một tổ chức và đưa ra định nghĩa tương ứng về sự cố của mô-đun AI, trong khi khoản 5.2.3 tóm tắt các mục đích và mục tiêu bao trùm cho cả AR và CR . Do đó, các tổ chức nên cân nhắc sử dụng tiêu chuẩn quản lý rủi ro liên quan đến tiêu chuẩn này. ISO 31000 [18] là một tiêu chuẩn thường được chấp nhận.

5.2 Yêu cầu và hướng dẫn về quản lý rủi ro

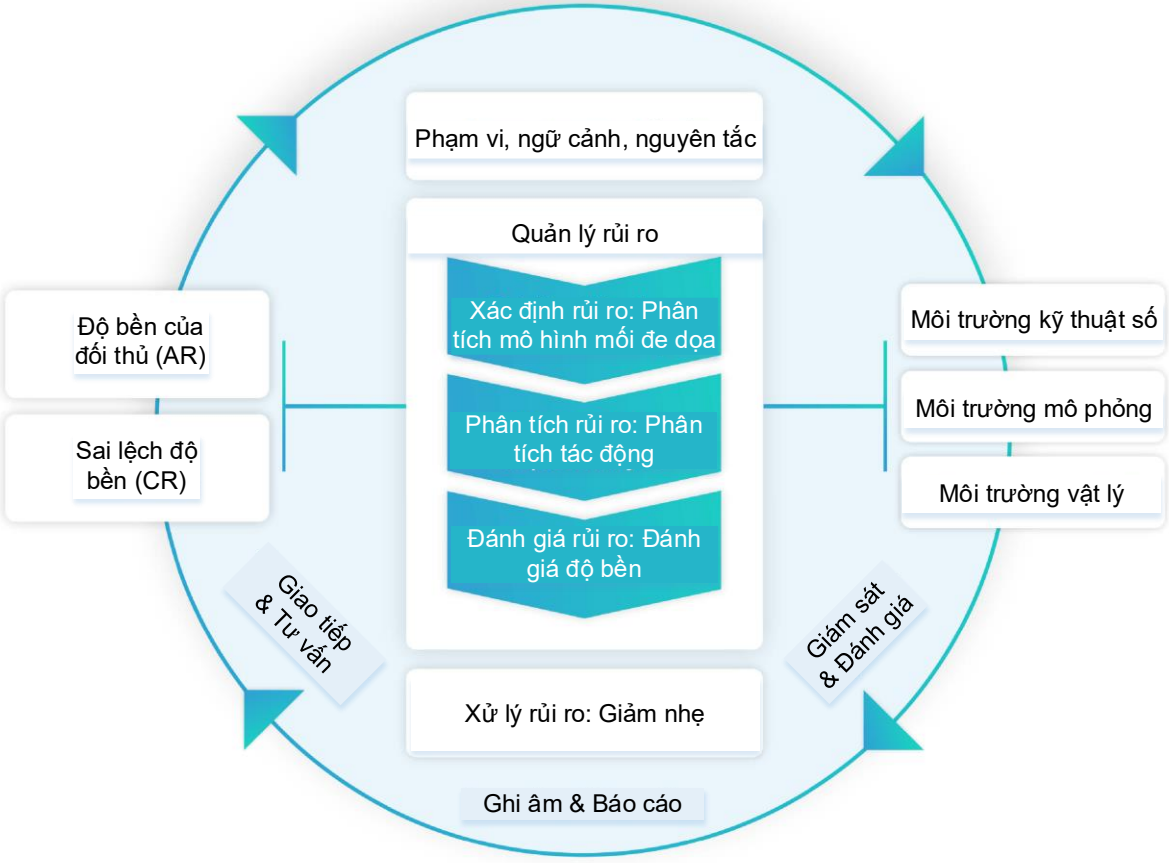
5.2.1 Tổng quan

Hình dưới đây (Hình 2) cung cấp hình ảnh trực quan về mối quan hệ giữa hai chiều của độ bền vững (AR và CR), các môi trường đánh giá khác nhau, tức là môi trường kỹ thuật số, mô phỏng và vật lý, đã được thúc đẩy trong phần trước và các giai đoạn của quá trình quản lý rủi ro.

Phía bên trái của Hình 2 nhấn mạnh rằng quy trình quản lý rủi ro AI được đề xuất khác nhau giữa AR và CR, tức là giữa bối cảnh ứng dụng mô-đun AI, nơi độ bền vững đối với nhiễu loạn của đối thủ được yêu cầu, và bối cảnh nơi nhiễu loạn do nguyên nhân tự nhiên là mối quan tâm. Do đó, các yêu cầu chất lượng AI cụ thể của tiêu chuẩn này cũng sẽ phản ánh và tương ứng với hai khía cạnh độ bền vững này. Ngoài sự khác biệt giữa AR và CR, các yêu cầu chất lượng dưới đây sẽ được cấu trúc theo các giai đoạn được mô tả của quy trình quản lý rủi ro, tức là Phạm vi, Bối cảnh, Tiêu chí, Đánh giá rủi ro và Xử lý rủi ro. Thiết kế quy trình quản lý rủi ro này phần lớn dựa trên quy trình quản lý rủi ro nổi tiếng của ISO 31000.

Giai đoạn Phạm vi, Bối cảnh và Tiêu chí hàm chứa các yêu cầu chất lượng giúp tùy chỉnh quy trình quản lý rủi ro tổng thể hơn phù hợp với tình hình cụ thể của tổ chức. Ở giai đoạn này, điều quan trọng là xác định tất cả các bên liên quan có liên quan và đặt ra các mục tiêu chung

bên ngoài liên quan đến AR và CR. Sau khi đặt ra những nền tảng tổ chức này, có thể tiến hành đánh giá rủi ro AR và CR. Giai đoạn này bao gồm ba giai đoạn phụ, gọi là Xác định rủi ro, Phân tích rủi ro và Đánh giá rủi ro. Do đó, các tổ chức phải tìm, thừa nhận và mô tả các rủi ro về độ bền vững gây nguy hiểm cho các mục tiêu độ bền vững chung được xác định trong Phạm vi, Bối cảnh và Tiêu chí. Trong trường hợp AR, tổ chức phải xác định các đối thủ tiềm ẩn cũng như các mục tiêu và ràng buộc liên quan của chúng ở giai đoạn này. Sau khi xác định rủi ro, các tổ chức sau đó chỉ định khả năng xảy ra và phân tích tác động tiêu cực tiềm ẩn của rủi ro đã xác định. Ở đây, điều cốt yếu là quyết định các phương pháp phù hợp để định lượng các kịch bản rủi ro khác nhau, và xác định phạm vi chấp nhận được đối với các biện pháp này. Điều này tạo điều kiện thuận lợi cho đánh giá rủi ro của mô hình AI quan tâm đối với các rủi ro đã được xác định của giai đoạn xác định rủi ro. Tùy thuộc vào kết quả đánh giá rủi ro, các biện pháp sau đó phải được thực hiện để cải thiện AR và/hoặc CR trong giai đoạn xử lý rủi ro.



Hình 2 – Quản lý rủi ro đối với AR và CR

Tất cả các giai đoạn của quy trình quản lý rủi ro đều được liên kết chặt chẽ với các môi trường có khả năng áp dụng của mô hình AI. Điều này được biểu thị ở phía bên bên phải của Hình 2. Ở đây, thuật ngữ Môi trường kỹ thuật số đề cập đến môi trường bao gồm dữ liệu đầu vào kỹ thuật số cho mô-đun AI. Do đó, nhiễu loạn trong môi trường kỹ thuật số bao gồm nhiễu loạn của dữ liệu đầu vào kỹ thuật số trong sạch. Các nhiễu loạn có thể thể hiện đặc tính của đối thủ hoặc có nguyên nhân "tự nhiên" (hoặc thậm chí bắt nguồn từ các mô hình của chúng). Chúng rõ ràng không bị hạn chế đối với những nhiễu loạn thủ công nhỏ. Chẳng hạn, các mô hình sương mù kỹ thuật số được áp dụng ngẫu nhiên với hình ảnh đầu vào kỹ thuật số cũng được bao trùm bởi định nghĩa này. Mặc dù những thay đổi này, theo quan điểm toán học, có thể có kích thước đáng kể xét theo chuẩn Lp (không giống như các nhiễu loạn của đối thủ), chúng không dễ thấy đối với người quan sát vì chúng tương ứng với các sự kiện xảy ra tự

nhiên.

Ngược lại, Môi trường mô phỏng bao gồm một quy trình tạo ra dữ liệu được thiết kế để đưa ra dữ liệu đầu vào tổng hợp cho mô-đun AI. Đối với môi trường mô phỏng, quá trình tạo ra dữ liệu hoặc các tham số của nó có thể có ý nghĩa đặc biệt. Nghĩa là, mô phỏng của dữ liệu đầu vào (ví dụ: cảnh đường phố dành cho lái xe tự động) bị nhiễu loạn. Chẳng hạn, các tư thế cụ thể hoặc đối tượng đối thủ "giả" có thể được đưa vào trong mô phỏng.

Cuối cùng, Môi trường vật lý là thế giới thực (tức là trước khi dữ liệu được chuyển dịch sang định dạng đầu vào kỹ thuật số thông qua cảm biến). Nhiễu loạn trong môi trường vật lý rõ ràng không được áp dụng cho đầu vào kỹ thuật số. Nhãn dán trên biển báo đường phố là một ví dụ phổ biến cho sự nhiễu loạn xảy ra trong miền vật lý.

Một số yêu cầu về chất lượng AI sau đây phải được giải quyết khác nhau đối với môi trường kỹ thuật số, mô phỏng hoặc vật lý. Ví dụ: các cuộc tấn công của đối thủ làm chuẩn nổi tiếng để đánh giá độ bền vững của mô hình AI, chẳng hạn như Phương pháp ký hiệu độ dốc nhanh [19] hoặc Hạ thấp độ dốc dự kiến [20], đã được giới thiệu để Đánh giá rủi ro của các mô hình AI trong môi trường kỹ thuật số. Do đó, các biện pháp khác (hoặc các bài kiểm tra căng thẳng) hoặc các điều chỉnh của các biện pháp truyền thống phải được xác định trong khi Phân tích rủi ro nếu tồn tại những quan ngại về môi trường vật lý của mô hình AI. Tuy nhiên, xin lưu ý rằng không phải cả ba môi trường đều phù hợp với mọi bối cảnh mô-đun AI và ứng dụng AI. Ví dụ, hãy xem xét một hệ thống đề xuất phim dựa trên AI. Một hệ thống như vậy hoạt động trong môi trường kỹ thuật số thuần túy và do đó không thể kiểm thử trong môi trường vật lý.

Khi nói đến việc đánh giá và cải thiện độ bền vững của một mô-đun AI, cũng có thể hữu ích khi xem xét cái gọi là kịch bản nhằm hướng dẫn người thử nghiệm trong suốt quá trình thử nghiệm và độc lập với môi trường triển khai và thử nghiệm của mô-đun AI. Một kịch bản đề cập đến các trường hợp hoặc chi tiết cụ thể xảy ra trong một thiết đặt hoặc môi trường nhất định (ví dụ: các điều kiện khí tượng cụ thể như mưa đá và băng giá). Ngoài ra, một kịch bản có thể được đặc trưng bởi một sự kiện hoặc chuỗi sự kiện cụ thể (ví dụ: một người đi bộ băng qua đường và dừng lại giữa chừng). Do đó, các kịch bản tạo thành các tập hợp con của tập dữ liệu gốc ở đó có thể nổi bật đặc trưng trong quy trình quản lý rủi ro (ví dụ: trong lái xe tự động, các chi tiết và đặc trưng của đô thị đặt ra những thách thức lớn hơn đáng kể về mặt an toàn và an ninh so với đường ở các vùng nông thôn dân cư thưa thớt).

Lưu ý rằng một kịch bản cần được lập mô hình/biểu diễn khác nhau trong mỗi môi trường thử nghiệm (một lần nữa, có thể là kỹ thuật số, mô phỏng hoặc có bản chất vật lý). Chẳng hạn, một kịch bản tương ứng với biển báo dừng xoay có thể được dịch sang môi trường kỹ thuật số bằng cách xoay các hình ảnh lành tính. Kịch bản tương tự có thể được mô hình hóa trong mô phỏng bằng cách thay đổi các tham số trong quá trình kết xuất theo cách mà các biển báo dừng được xoay theo mặc định. Trong môi trường vật chất, biển báo dừng sẽ trải qua các vòng quay vật lý thực tế, cuối cùng thu được dữ liệu ảnh.

Hơn nữa, cần phải nhấn mạnh rằng quản lý rủi ro độ bền vững AI không thể được xem như một quy trình cứng nhắc chỉ được thực hiện một lần cho mọi mô hình AI. Thay vào đó, cần phải liên tục làm việc thông qua các giai đoạn khác nhau của quy trình quản lý rủi ro. Các mô hình mối đe dọa được phát hiện và rủi ro về độ bền vững có thể thay đổi thường xuyên do những thay đổi trong môi trường hoặc sự xuất hiện của các đối thủ mạnh hơn. Công việc liên tục cần thiết này thông qua quy trình quản lý rủi ro được quan sát bằng vòng tròn màu xanh bên dưới trong Hình 2.

Trong khoản con tiếp theo, các quy trình quản lý rủi ro được liên kết với các mục tiêu tổng thể của một tổ chức và các nhu cầu về an ninh và an toàn của nó. Các yêu cầu cơ bản này xác định lỗi mô-đun AI và kiến thức về các tình huống quan trọng, bộ dữ liệu, cũng như môi trường

mô phỏng và thử nghiệm. Để tạo điều kiện thuận lợi cho việc tham gia với các yêu cầu được trình bày trong các điều 5.3 và 5.4 và việc sử dụng chúng trong chiến lược tổng thể của tổ chức, các mục tiêu và mục đích chung liên quan đến AR và CR sẽ được thể thức hóa. Các mục tiêu và mục đích chung phản ánh dòng lập luận trong các điều 5.3 và 5.4, tức là chúng được phát triển liên quan đến phân tích mô hình mối đe dọa, phân tích khả năng xảy ra & tác động, đánh giá độ bền vững cũng như các biện pháp giảm thiểu.

Để tạo thuận lợi cho công việc với các yêu cầu chất lượng AI, tất cả các yêu cầu độ bền vững được giới thiệu trong Thông số kỹ thuật DIN này đều được cung cấp một cách có hệ thống với mã định danh văn bản có dạng Rxx_yyy.n.

“R” là viết tắt của độ bền vững, “xx” đề cập đến loại tổng quát, “yyy” là loại phụ và “n” là số thứ tự trong loại phụ. Trong tiêu chuẩn này, ba loại được sử dụng: "GE" cho độ bền vững chung, "AR" cho độ bền vững trước đối thủ và "CR" cho sai lệch độ bền vững. Ví dụ: “RAR_TMA.1” xác định yêu cầu đầu tiên liên quan đến "Phân tích mô hình mối đe dọa" (TMA) theo loại “AR”.

5.2.2 Phạm vi, bối cảnh và tiêu chí

RGE_SCC.1 Thiết lập một liên kết giữa các hoạt động quản lý của một tổ chức và quản lý rủi ro liên quan đến AI. Khi quản lý rủi ro AI, các tổ chức nên phân tích phạm vi của hoạt động quản lý cũng như bối cảnh bên trong và bên ngoài của chúng. Liên quan đến độ bền vững, các tổ chức nên cung cấp các tiêu chí chung để đáp ứng mong đợi của các bên liên quan. ISO 31000 [18] cung cấp hướng dẫn toàn diện về những vấn đề này.

CHÚ THÍCH: Các mục tiêu được mô tả trong 5.2.3.1 và 5.2.3.2 hỗ trợ quá trình xác định rủi ro và phân tích rủi ro được mô tả trong ISO 31000 [8]. Đánh giá rủi ro được hỗ trợ bởi các yêu cầu quy định trong 5.2.3.3. Quá trình xử lý rủi ro được hỗ trợ bởi 5.2.3.4.

CHÚ THÍCH 2 ISO 31000 [18] định nghĩa “rủi ro” là “ảnh hưởng của sự không chắc chắn đối với các mục tiêu” và thừa nhận rằng những ảnh hưởng như vậy có thể là tiêu cực cũng như tích cực. T êu chuẩn này, xử lý đặc biệt với các sự kiện và tình huống gây nguy hiểm cho độ bền vững của hệ thống AI, tự hạn chế các tác động dẫn đến kết quả tiêu cực. Do đó, xử lý rủi ro được giới hạn trong việc xác định và thực hiện các kiểm soát và biện pháp đối phó giảm thiểu.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, dữ liệu, nền tảng, môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển, triển khai, vận hành, ngừng hoạt động
Quy trình/các quy trình vòng đời	Quy trình quản lý nguồn nhân lực, Quy trình quản lý chất lượng, Quy trình quản lý nhận thức, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đảm bảo chất lượng, Quy trình xác định nhu cầu và yêu cầu của các bên liên quan, Quy trình xác định yêu cầu hệ thống/phần mềm

RGE_SCC.2 Dựa trên các mục tiêu của tổ chức và nhu cầu về an toàn và bảo mật, cần phải xem xét quyết định và tài liệu hóa những nhiệm vụ được AI hỗ trợ trong quản lý rủi ro liên quan đến độ bền vững. Đối với các nhiệm vụ mà AI hỗ trợ được coi là có liên quan, cung cấp một giải thích chi tiết về môi trường, nền tảng, dữ liệu và các mô hình. Điều này bao gồm một mô tả rõ ràng về nhiệm vụ máy học mà mô-đun AI đã được huấn luyện đối với các tác động về an toàn và bảo mật tiềm ẩn. Mô tả kiến trúc, bao gồm các tiện ích bổ sung cho mô-đun, cũng như các khả năng và hiệu năng thử nghiệm của nó.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, dữ liệu, nền tảng, môi trường

Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển, triển khai, vận hành, ngừng hoạt động
Quy trình/các quy trình vòng đời	Quy trình quản lý nguồn nhân lực, Quy trình quản lý chất lượng, Quy trình quản lý nhận thức, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đảm bảo chất lượng, Quy trình xác định nhu cầu và yêu cầu của các bên liên quan, Quy trình xác định yêu cầu hệ thống/phần mềm

RGE_SCC.3 Xác định lỗi mô-đun AI. Trong mọi trường hợp, dẫn suất ra một định nghĩa cụ thể, định lượng và khả thi về "lỗi mô-đun AI". Điều này tương ứng với việc thiết lập một danh mục các số liệu và thước đo cũng như chỉ định các giá trị hoặc phạm vi quan trọng. Cả định nghĩa dẫn xuất và danh mục các số liệu và thước đo phải đủ sắc thái để tính toán cho các nguyên nhân và loại lỗi hệ thống khác nhau. Điều này liên quan đến phân tích kỹ lưỡng về tiềm năng phân nhánh của lỗi hệ thống.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, dữ liệu, nền tảng, môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển, triển khai, vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý mô hình vòng đời, Quy trình quản lý chất lượng, Quy trình quản lý nhận thức, Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý cấu hình, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình đảm bảo chất lượng, Quy trình xác định yêu cầu hệ thống/phần mềm, Quy trình phân tích hệ thống, Quy trình xác minh

RGE_SCC.4 Chỉ định các kịch bản quan trọng và xác định các tập dữ liệu liên quan, môi trường mô phỏng cũng như môi trường thử nghiệm trong thế giới thực sẽ được xem xét trong khi đánh giá rủi ro. Đối với một nhiệm vụ máy học nhất định, xác định các tình huống, đó là bảo mật hoặc an toàn quan trọng, và/hoặc thường gặp phải trong quá trình triển khai. Đối với các kịch bản này, cần ghi lại các tập dữ liệu có liên quan, môi trường mô phỏng cũng như môi trường thử nghiệm trong thế giới thực tiềm năng và đưa ra lập luận về lý do tại sao dữ liệu, mô phỏng và môi trường thử nghiệm trong thế giới thực đã định lại là đại diện cho các kịch bản được coi là cực kỳ quan trọng, đồng thời cũng giải quyết những hạn chế có thể xảy ra. Ngoài ra hoặc theo cách khác, để tăng thêm trọng lượng cho các kịch bản cụ thể, bất kể chúng có gây ra rủi ro về mặt an toàn và bảo mật hay không, các vùng nhất định của tập dữ liệu có thể phải được ưu tiên, cho phép người dùng phát triển các chiến lược quản lý rủi ro tinh vi được cân bằng tốt.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, dữ liệu, môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình đảm bảo chất lượng, Quy trình phân tích hệ thống

5.2.3 Mục tiêu và Mục đích chung

5.2.3.1 Phân tích mô hình mối đe dọa

Mục tiêu 1 Đặc trưng của các đối thủ và nhiễu loạn của đối thủ. Xem xét AR, các đối thủ có thể có và kiến thức tiềm năng của họ cũng như quyền truy cập vào mô-đun AI nhất định sẽ

được xác định và định rõ đặc điểm. Hơn nữa, các định nghĩa về các loại mẫu của đối thủ khác nhau (hoặc nhiễu loạn của đối thủ) trong ngữ cảnh ứng dụng nhất định sẽ được cung cấp.

Mục tiêu 2 Đặc trưng cho các loại sai lệch hoặc thay đổi phân phối. Xem xét CR, có thể có các nguồn sai lệch hoặc thay đổi phân phối trong bối cảnh ứng dụng nhất định sẽ được xác định, nghĩa là nó cần phải đánh giá, đo lường và tài liệu hóa các cách thức mà dữ liệu gặp phải trong khi triển khai có thể khác với các tập huấn luyện được thu thập trước đó (theo cách tự nhiên, nghĩa là không có sự can thiệp trên một phần của kẻ tấn công độc hại).

5.2.3.2 Phân tích khả năng xảy ra và tác động

Mục tiêu 1 Đánh giá hậu quả của việc thiếu độ bền vững. Các hậu quả tiềm ẩn của lỗi mô-đun AI sẽ được đánh giá để phân tích mối liên quan của cả AR và CR. Điều này liên quan đến việc đánh giá khả năng hợp lý và kỹ lưỡng về khả năng xảy ra và những rủi ro của từng phân nhánh tiềm năng của lỗi hệ thống. Mọi phân nhánh có thể xảy ra sẽ được tài liệu hóa cùng với khả năng xảy ra và những thiệt hại dự kiến. Hơn nữa, các mô-đun AI phải được phân loại rõ ràng thành rủi ro cao hoặc thấp.

5.2.3.3 Đánh giá độ bền vững

Mục tiêu 1 Lặp lại đánh giá AR sử dụng dữ liệu đối thủ cũng như các số liệu và đo lường thích hợp. Nếu AR với các loại mẫu của đối thủ chắc chắn được xác định là một thách thức đáng kể, thì độ bền vững của mô-đun AI đối với các mẫu của đối thủ có liên quan sẽ được đánh giá. Các kỹ thuật giảm thiểu tiềm năng và biện pháp đối phó cũng sẽ được đánh giá sau khi chúng được thực hiện. Tất cả các thử nghiệm phải tuân theo một quy trình được quy định bởi một kế hoạch thử nghiệm chỉ định các phạm vi, khoảng thời gian và kiểm chuẩn có thể chấp nhận được của các số liệu và đo lường phù hợp theo quy định của các bên liên quan.

Mục tiêu 2 Lặp lại đánh giá CR sử dụng dữ liệu bị hỏng cũng như các số liệu và đo lường thích hợp. Đối với CR, cần phải lựa chọn và tài liệu hóa các số liệu và đo lường để xác định loại (cũng như mức độ nghiêm trọng, phạm vi và khoảng thời gian) của tsai lệch hoặc thay đổi phân phối ở đó mô-đun AI có thể chịu đựng được mà không làm suy giảm hiệu năng. Bằng các số liệu và đo lường đánh giá phù hợp, cũng như dữ liệu thể hiện các loại sai lệch nghiêm trọng hoặc thay đổi phân phối đã xác định, mô hình đã huấn luyện sẽ được kiểm tra và so sánh. Các loại và mức độ dẫn đến suy giảm hiệu năng phải được lập thành văn bản.

5.2.3.4 Sự giảm nhẹ

Mục tiêu 1 Xác định và phát triển một tập hợp các chiến lược phòng thủ để đạt được AR. Đối với AR, các phòng thủ thích hợp sẽ được lựa chọn, thực thi và thử nghiệm chống lại các cuộc tấn công của đối thủ phù hợp và các số liệu được xác định trước khác. Trong quá trình phát triển, triển khai và vận hành, nó sẽ được đánh giá liên tục xem có cần thiết phải phòng thủ thêm hay không.

Mục tiêu 2 Xác định và phát triển một tập hợp các kỹ thuật giảm thiểu để đạt được CR. Đối với CR, một tập hợp các chiến lược sẽ được xác định để giảm thiểu tác động của thay đổi phân phối hoặc sai lệch đã được phát hiện. Ngoài ra, một tập hợp các biện pháp đối phó sẽ được xác định, sẵn sàng được sử dụng trong trường hợp phát hiện thấy sự thay đổi phân phối trong thời gian chạy.

5.3 Yêu cầu cụ thể với độ bền vững trước đối thủ

5.3.1 Phạm vi, bối cảnh và tiêu chí

Các yêu cầu được liệt kê trong 5.2.2 hoàn toàn có thể áp dụng cho các cân nhắc về độ bền vững trước đối thủ.

5.3.2 Phân tích mô hình mối đe dọa

5.3.2.1 RAR_TMA.1

Xác định các đặc điểm của đối thủ tiềm ẩn và phân lớp các loại nhiễu loạn của đối thủ mà chúng có thể tạo ra. Để đánh giá mối liên quan về độ bền vững trước đối thủ đối với từng mô-đun AI cụ thể, các đối thủ tiềm ẩn phải được đặc tả. Điều này bao hàm một tài liệu tỉ mỉ về các nhiễu loạn của đối thủ mà chúng có thể tạo ra trong một kịch bản nhất định, với một kịch bản cụ thể yêu cầu một tập hợp riêng biệt hoặc loại nhiễu loạn của đối thủ. Việc xác định và phân tích các kịch bản bảo mật quan trọng tiềm ẩn được xây dựng dựa trên 5.2.2 RGE_SCC.4.

VÍ DỤ: Các đối thủ có liên quan có thể khác nhau về mục tiêu của chúng (ví dụ: tấn công có mục tiêu so với không có mục tiêu, phạm vi tấn công riêng lẻ so với phổ quát), kiến thức về mô hình nạn nhân (ví dụ: thiết đặt hộp trắng so với hộp đen) và các thuộc tính tương tự.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, nền tảng, môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển, vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình đảm bảo chất lượng, Quy trình xác định yêu cầu hệ thống/phần mềm, Quy trình phân tích hệ thống

5.3.2.2 RAR_TMA.2

Chỉ định cách đối thủ đạt được sự không rõ ràng của nhiễu loạn trong ngữ cảnh ứng dụng cụ thể. Một tiền đề quan trọng về một mẫu của đối thủ là nó phải không thể nhận thấy hoặc ít nhất là không dễ thấy đối với mắt người. Một mẫu của đối thủ không dễ thấy có thể được định nghĩa theo nhiều cách khác nhau. Các loại đầu vào không thể nhận thấy hoặc không dễ thấy khác nhau phải được đặc tả và đặt trong mối quan hệ với các loại truy cập đầu vào mà đối thủ đã được phát hiện hoặc được giả định là có.

VÍ DỤ 1: Theo giả định về môi trường kỹ thuật số trong thị giác máy tính, đối thủ có thể quyết định chỉ tấn công một vài pixel để làm cho cuộc tấn công của chúng không thể nhận thấy được. Trong một môi trường mô phỏng, các đối thủ sẽ có khả năng đặt các đối tượng của đối thủ vào cảnh, xoay nhẹ các biển báo trên đường hoặc tìm vị trí chính xác của người đi bộ, khiến cho mô-đun AI phân loại dữ liệu sai.

VÍ DỤ 2: Trong miền xử lý ngôn ngữ tự nhiên, các từ đồng nghĩa được chọn phù hợp có thể che giấu một cuộc tấn công của đối thủ theo cách mà cảm xúc của văn bản bị xáo trộn vẫn giữ nguyên đối với người quan sát, nhưng sẽ được mô hình phân loại văn bản đánh dấu khác biệt. Ngoài ra, toàn bộ tài liệu có thể được diễn giải mang lại kết quả có hại tương tự mà người đọc không chú ý.

Lớp yêu cầu	Được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Dữ liệu, môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển, vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình đảm bảo chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm

5.3.2.3 RAR_TMA.3

Xác định sự thành công về nhiễu loại của đối thủ. Xây dựng trên định nghĩa về lỗi mô-đun AI (xem 5.2.2 RGE_SCC.3), cung cấp một định nghĩa phù hợp (có thể là định nghĩa hoạt động

dựa trên các giả định đầu tiên) về một cuộc tấn công "thành công" vào mô-đun AI, tức là một mẫu của đối thủ. Điều này phải được thực hiện tương ứng với các định nghĩa về đối thủ tiềm ẩn và tính không dễ thấy của đầu vào như được mô tả trong RAR_TMA.1 và RAR_TMA.2 tương ứng.

VÍ DỤ: Vì các mẫu của đối thủ ban đầu được xác định trong bối cảnh của các hệ thống phân loại cho thị giác máy tính, nên định nghĩa về một mẫu của đối thủ không phải luôn rõ ràng. Ví dụ: đối với nhiệm vụ phân đoạn ngữ nghĩa, có thể sẽ không đủ nếu một pixel đơn lẻ thay đổi phân loại của nó do nhiễu loạn của đối thủ.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình đảm bảo chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm, Quy trình thực hiện

5.3.2.4 RAR_TMA.4

Xác định phạm vi nhiễu loạn của đối thủ, tức là hoặc các đối thủ tiềm ẩn có nhắm đến tạo ra nhiễu loạn riêng biệt hoặc nhiễu loạn toàn cục. Chuẩn bị với những đối thủ có ý định hoặc có khả năng tạo ra nhiễu loạn phổ quát của đối thủ, tức là nhiễu loạn mà ở đó sự thành công trên hoặc là một phạm vi các kịch bản khác nhau hoặc là như một sự lựa chọn, tất cả các điểm dữ liệu gắn liền với một kịch bản cụ thể (xem 5.2.2 RGE_SCC.3).

VÍ DỤ: Một nhiễu loạn của đối thủ có thể áp dụng phổ biến hoặc rộng rãi nếu nó thành công trên một phạm vi các phối cảnh máy ảnh hoặc dưới các điều kiện ánh sáng khác nhau trong một nhiệm vụ thị giác máy tính hoặc, trong lĩnh vực NLP, trong nhiều bối cảnh ngữ nghĩa hoặc thực dụng.

Lớp yêu cầu	Được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình kế hoạch hóa đề án, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đảm bảo chất lượng

5.3.2.5 RAR_TMA.5

Chỉ định cách phân biệt các đối thủ và các loại nhiễu loạn của đối thủ có thể được thể hiện hoặc mô phỏng trong các môi trường thử nghiệm khác nhau. Dựa trên đặc điểm của các đối thủ tiềm ẩn trong RAR_TMA.1, trong đó bao hàm tính toán kỹ lưỡng khả năng của chúng, mô tả cách các đối thủ này có thể được biểu hiện hoặc mô phỏng trong các môi trường thử nghiệm khác nhau (xem 5.2.2 RGE_SCC.4), chúng là vật lý, kỹ thuật số, hoặc mô phỏng. Đối với từng môi trường thử nghiệm dưới sự xem xét, sửa đổi và diễn giải lại các quan điểm về tính không dễ thấy (xem RAR_TMA.2), sự thành công (xem RAR_TMA.3) và phạm vi (xem RAR_TMA.4) nhiễu loạn của đối thủ.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đảm bảo chất lượng, Quy trình phân tích hệ thống

5.3.2.6 RAR_TMA.6

Hợp nhất các cân nhắc ở trên (xem RAR_TMA.1, RAR_TMA.2, RAR_TMA.3, RAR_TMA.4 và RAR_TMA.5) thành danh mục các đối thủ tiềm ẩn, các loại nhiễu loạn, các định nghĩa thành công, và phạm vi của từng nhiễu loạn, nhận thức các kịch bản và môi trường thử nghiệm khác nhau. Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu (được gọi là "đặc tính AR") có chứa phân tích cấp cao về mô hình mối đe dọa, tức là đặc điểm của đối thủ và các loại nhiễu loạn của đối thủ tương ứng mà chúng có thể tạo ra. Đặc tính AR phải luôn phân biệt giữa các kịch bản và môi trường thử nghiệm khác nhau. Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình xác định các yêu cầu hwj thống/phần mềm

5.3.3 Phân tích Khả năng xảy ra & Tác động

5.3.3.1 RAR_LIA.1

Đánh giá hậu quả tiềm ẩn của một cuộc tấn công thành công, dựa trên các định nghĩa trong 5.3.2. Một phân tích trong lĩnh vực này nên được cấu trúc xung quanh các kịch bản được coi là quan trọng như được trình bày trong 5.2.2 RGE_SCC.4, khả năng của đối thủ (xem 5.3.2 RAR_TMA.1 và RAR_TMA.2), và các loại nhiễu loạn có thể xảy ra của đối thủ. Do đó, về cơ bản sử dụng đặc tính AR như được thiết lập trong 5.3.2 RAR_TMA.6, phân tích sẽ xác định loại hành vi không mong muốn hoặc thậm chí có hại mà mô-đun AI đã định có thể biểu hiện (xem 5.2.2 RGE_SCC.3) như một kết quả hành động của một đối thủ.

CHÚ THÍCH: Ở đây chỉ quan tâm đến các hậu quả ở cấp độ mô-đun AI, các yêu cầu về an toàn và/hoặc bảo mật từ cấp độ hệ thống phải được chuyển thành các yêu cầu độ bền vững trước đối thủ mà mô-đun AI phải được đáp ứng (xem 5.2.2).

VÍ DỤ: Các hậu quả tiềm ẩn đã được mô tả trong [3] và được đặc trưng bởi sự vi phạm tính toàn vẹn, tính khả dụng, tính bí mật hoặc quyền riêng tư. Vi phạm tính toàn vẹn tương ứng với đầu ra của mô-đun AI bị xâm phạm, điều này có thể dẫn đến suy giảm giá trị độ tin cậy hoặc thậm chí phân loại sai. Vi phạm tính khả dụng có thể dẫn đến suy giảm quyền truy cập vào mô-đun AI hoặc giảm tốc độ suy luận. Vi phạm bí mật cung cấp cho kẻ tấn tiềm ẩn công kiến thức về mô-đun AI hoặc dữ liệu mà nó đã được huấn luyện. Cuối cùng nhưng không kém phần quan trọng, vi phạm quyền riêng tư là một trường hợp đặc biệt của vi phạm bí mật khi kẻ tấn công lại hiểu thấu được thông tin cá nhân có trong dữ liệu huấn luyện, vi phạm luật bảo vệ dữ liệu.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình phân tích hệ thống, Quy trình xác minh

5.3.3.2 RAR_LIA.2

Phân tích tính khả thi về nhiễu loạn của đối thủ và rút ra ước tính khả năng xảy ra cho từng loại nhiễu loạn của đối thủ, và nói chung hơn là sự tồn tại của các đối thủ với các tập hợp khả năng khác nhau. Một phân tích khả thi kỹ lưỡng về các loại nhiễu loạn khác nhau của đối thủ (xem 5.3.2 RAR_TMA.6) giúp đánh giá mức độ phù hợp của độ bền vững trước đối thủ. Phân tích tính khả thi phải được liên kết với đánh giá như thế nào mmỗi loại nhiễu loạn của đối thủ sẽ xảy ra trong bối cảnh một đối thủ sử dụng một tập hợp các khả năng cụ thể. Tuy nhiên, với các kịch bản khác nhau đã định (xem 5.2.2 RGE_SCC.4), đánh giá có thể mang lại các ước tính khả năng xảy ra khác nhau cho từng loại nhiễu loạn của đối thủ.

VÍ DỤ: Ví dụ: nếu mô-đun AI chỉ bị đe dọa bởi các mẫu của đối thủ hộp trắng, phổ biến, vật lý, đắt tiền về mặt tính toán, thì có thể không tồn tại mối nguy hiểm thực tế và do đó, quan điểm và phát triển các chiến lược phòng thủ cho loại tấn công của đối thủ cụ thể này có thể không cần thiết

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm, Quy trình phân tích hệ thống

5.3.3.3 RAR_LIA.3

Quyết định có liên quan chung về độ bền vững trước đối thủ bằng cách thực hiện phân loại rủi ro của các đối thủ khác nhau và các loại nhiễu loạn của đối thủ. Nó phải được lập thành tài liệu và giải thích, liệu độ bền vững trước những nhiễu loạn của đối thủ là một thành phần chất lượng quan trọng đối với mô-đun AI có được xem xét hay không. Điều này liên quan đến việc thực hiện phân loại rủi ro của các loại nhiễu loạn của đối thủ được xác định trước bằng cách sử dụng tiêu chí rủi ro được mô tả trong 5.2 Yêu cầu và Hướng dẫn về Quản lý Rủi ro (cụ thể là 5.2.2 Phạm vi, Bối cảnh và Tiêu chí), trong đó rủi ro được khái niệm hóa như là sản phẩm thiệt hại của một cuộc tấn công của đối thủ (xem RAR_LIA.1) và khả năng xảy ra của nó (xem RAR_LIA.2).

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu

Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.4 RAR_LIA.4

Xác định và các kịch bản ưu tiên trong đó độ bền vững trước đối thủ là cần thiết. Có thể không cần thiết, ngăn cản khả năng đơn độc đạt được đầy đủ độ bền vững trước đối thủ trong mọi trường hợp. Nói cách khác, các kịch bản xác định và ưu tiên trong đó độ bền vững trước đối thủ là không bỏ qua được (xem 5.2.2 RGE_SCC.4).

VÍ DỤ: Các kịch bản thuộc loại này tương ứng với các vùng của tập dữ liệu với một số siêu thông tin cụ thể. Ví dụ, trong lĩnh vực máy bay không người lái, bối cảnh đô thị, đông dân cư đòi hỏi phải cẩn thận hơn so với bối cảnh ở nông thôn mở.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.5 RAR_LIA.5

Thiết lập tài liệu bao hàm các cuộc tấn công, đo lường và số liệu định lượng độ bền vững với các đối thủ đã xác định và các loại nhiễu loạn của đối thủ. Xác định và thu thập các cuộc tấn công, đo lường và số liệu định lượng, hoặc ít nhất là đưa ra dấu hiệu về độ bền vững liên quan đến các đối thủ và các loại nhiễu loạn của đối thủ (xem 5.3.2 RAR_TMA.6). Chỉ định các định nghĩa và khái niệm toán học hoặc thuật toán liên quan tới các kịch bản và môi trường thử nghiệm (xem 5.2.2 RGE_SCC.4), cung cấp cấu trúc của tài liệu hoặc danh mục đã nói trên. Sau đây, danh mục này sẽ được gọi là "danh mục đo lường và tấn công của đối thủ".

VÍ DỤ: Một cách tiếp cận chung để đánh giá độ nhạy của mô-đun AI đối với một mô hình mối đe dọa cụ thể là tạo ra một tập hợp các cuộc tấn công có liên quan của đối thủ phản ánh một đối thủ cụ thể và loại nhiễu loạn tương ứng, với tập hợp các cuộc tấn công tạo thành một phần tích hợp danh mục tấn công và đo lường đối thủ. Bên cạnh các cuộc tấn công, đã phải xác định các số liệu ở đó có thể tổng hợp các kết quả tấn công và đưa ra tuyên bố chung về độ nhạy đối với mô hình mối đe dọa nhất định. Mặc dù trong trường hợp chung, các số liệu này có thể dễ dàng dẫn xuất từ việc xác định một cuộc tấn công thành công, nhưng nó là tiêu chuẩn thực hành để tính toán độ chính xác tổng thể của đối thủ và so sánh nó với độ chính xác rõ ràng.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành

Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.6 RAR_LIA.6

Tích hợp quy trình cập nhật thường xuyên cho danh mục đo lường và tấn công của đối thủ (xem RAR_LIA.5), tức là thường xuyên tiến hành đánh giá trạng thái của nghệ thuật. Do tốc độ nghiên cứu nhanh chóng trong lĩnh vực này, danh mục tấn công và đo lường đối thủ phải được cập nhật thường xuyên.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm, Quy trình vận hành, Quy trình bảo dưỡng

5.3.3.7 RAR_LIA.7

Giữ cho danh mục đo lường và tấn công của đối thủ (xem RAR_LIA.5) được cân bằng. Nếu nó là khả thi về mặt tính toán, cũng sử dụng các chỉ số không dựa trên tấn công để xác định lỗ hổng của mô-đun AI với các loại khác nhau của đối thủ và nhiễu loạn của đối thủ. Ngoài ra, sử dụng các số liệu tổng quát để đưa ra các tuyên bố khoảng cách chung hơn hoặc đảm bảo độ bền vững.

CHÚ THÍCH: Trước đây, độ bền vững chủ yếu được đánh giá sơ bộ bằng các cuộc tấn công của đối thủ, ví dụ: bằng cách so sánh độ chính xác rõ ràng với độ chính xác thu được trong các điều kiện của đối thủ. Thật không may, các đo lường tổng hợp này thường tạo ra cảm giác sai lệch về độ nhạy tổng thể của mô-đun AI đối với các nhiễu loạn của đối thủ. Vì vậy, sử dụng thêm, cách tiếp cận bổ sung là rất thích hợp.

VÍ DỤ: Các kỹ thuật xác minh chính thức có thể được sử dụng để đưa ra các đảm bảo độ bền vững đối với đặc tính mô hình mối đe dọa đã định. Tuy nhiên, chúng không mở rộng tốt đối với các vấn đề thị giác máy tính nặng về tính toán.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.8 RAR_LIA.8

Lựa chọn đo lường kiểm chuẩn thích hợp cho các số liệu và phép đo có liên quan, tức là kết nối danh mục đo lường và tấn công của đối thủ (xem RAR_LIA.5) với các giá trị có khả năng chấp nhận được hoặc mong muốn. Ưu tiên các đối thủ đã xác định và các loại nhiễu loạn của đối thủ (xem 5.3.2 RAR_TMA.1 và RAR_TMA.2), xem xét, lựa chọn và tổng hợp các đo lường kiểm chuẩn cho đo lường và số liệu của đối thủ liên quan đến kịch bản và môi trường thử

nghiệm cụ thể (xem 5.2.2 RGE_SCC.4). Điều này là cần thiết để tránh sai lệch trong đánh giá kết quả.

LƯU Ý: Nếu có khả năng, hãy so sánh với các tiêu chuẩn công nghiệp hoặc quy định.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.9 RAR_LIA.9

Lựa chọn và tài liệu hóa phạm vi và khoảng thời gian có khả năng chấp nhận phù hợp để đo lường và số liệu đã định. Sử dụng các đo lường kiểm chuẩn (xem RAR_LIA.8) để xác định hoặc rút ra các khoảng thời gian và phạm vi thích hợp để đo lường và số liệu của danh mục đo lường và tấn công của đối thủ mà mô-đun AI cần đáp ứng để nó có thể được xem như bền vững với các đối thủ đã xác định và các loại nhiễu loạn của đối thủ.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.3.10 RAR_LIA.10

Hợp nhất các cân nhắc ở trên (xem RAR_LIA.1, RAR_LIA.2, RAR_LIA.3, RAR_LIA.4 và RAR_LIA.5) bằng cách mở rộng tài liệu được gọi là “đặc tính AR”. Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu (được gọi là "đặc tính AR mở rộng (phân tích khả năng xảy ra và tác động)"). Tiêu chuẩn này bao gồm phân tích các hậu quả tiềm ẩn và khả năng xảy ra lỗi mô-đun AI do thiếu độ bền vững đối với loại nhiễu loạn của đối thủ nhất định, cũng như tương ứng với phân loại rủi ro. Đặc tính AR mở rộng (phân tích khả năng xảy ra và tác động) hơn nữa còn bao gồm danh mục các kịch bản bảo mật quan trọng được ưu tiên (xem 5.2.2 RGE_SCC.4) cũng như danh mục đo lường và tấn công của đối thủ (xem RAR_LIA.5). Tiêu chuẩn phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trinh xác định các yêu cầu hệ thống/phần mềm

5.3.4 Đánh giá độ bền vững

5.3.4.1 RAR_EVL.1

Phát triển một kế hoạch thử nghiệm (AR) dựa trên đặc tính AR mở rộng (phân tích khả năng xảy ra và tác động) (xem 5.3.3 RAR_LIA.10). Trong kế hoạch thử nghiệm, hãy tính đến các khía cạnh sau, đồng thời luôn cung cấp các luận cứ và tài liệu cho các lựa chọn của bạn:

- Kế hoạch thử nghiệm phải chứa các tiêu chí đưa vào phù hợp, tức là các sự kiện phải chỉ định kích hoạt áp dụng kế hoạch thử nghiệm.

- Kế hoạch thử nghiệm phải bao gồm các tiêu chí để dừng sớm, nghĩa là các sự kiện hủy bỏ thử nghiệm phải được mô tả.

- Các kết quả đánh giá phải được tiêu chuẩn hóa, hoặc ít nhất là có thể so sánh được, để cho phép đưa ra các kết luận rõ ràng, với kế hoạch thử nghiệm hướng dẫn cam kết quan trọng với các kết quả thu được.

- Kế hoạch thử nghiệm phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình kế hoạch đề án, Quy trình quản lý quyết định, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm, Quy trình xác minh

5.3.4.2 RAR_EVL.2

Tích hợp quy trình cập nhật thường xuyên cho kế hoạch thử nghiệm (xem RAR_EVL.1), tức là thường xuyên tiến hành đánh giá trình độ phát triển và quyết định xem có thay đổi cơ bản nào xảy ra hay không (liên quan đến loại thay đổi phải được xác định trước bởi người dùng). Do tốc độ nghiên cứu nhanh chóng trong lĩnh vực này, nên danh mục tấn công và đo lường đối thủ phải được cập nhật sai lệch thường xuyên. Độ bền vững trước đối thủ của mô-đun AI phải được đánh giá lại và thử nghiệm lại sau mỗi lần cập nhật danh mục tấn công và đo lường đối thủ hoặc kế hoạch thử nghiệm.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm, Quy trình vận hành, Quy trình bảo trì

5.3.4.3 RAR_EVL.3

Sau mỗi lần chạy kế hoạch thử nghiệm (xem RAR_EVL.1), hãy so sánh kết quả với phạm vi có khả năng chấp nhận được và kết quả kiểm chuẩn (xem 5.3.3 RAR_LIA.5) và quyết định

xem mô-đun AI có đủ bền vững hay không. Mọi vi phạm kiểm chuẩn phải dẫn đến xem xét lại chiến lược phòng thủ và dừng sơ bộ, thậm chí là cân nhắc lại việc triển khai mô-đun AI.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trinh xác định các yêu cầu hệ thống/phần mềm, Quy trình vận hành, Quy trình bảo trì

5.3.4.4 RAR_EVL.4

Hợp nhất các kết quả và thông tin ở trên (xem RAR_EVL.1, RAR_EVL.2 và RAR_EVL.3) dưới dạng tài liệu có cấu trúc (được gọi là "Đánh giá AR"), tức là thu thập và lưu trữ kỹ lưỡng các kết quả đánh giá và tất cả thông tin về hoạt động và cập nhật kế hoạch thử nghiệm cũng như dữ liệu đối thủ được tạo ra để có thể được sao chép sau này. Các phần thông tin phù hợp nhất về việc thực hiện và cập nhật kế hoạch thử nghiệm phải được thu thập và lưu trữ (ví dụ: thống kê tóm tắt). Ngoài ra, dữ liệu đối thủ được phát hiện phải được giảm xuống thành một bản tóm tắt toàn diện về cả tạo ra và phân nhánh của chúng, với khả năng tái tạo cấu thành mục tiêu chính (ví dụ: đồ tạo tác đã lưu). Nói cách khác, thông tin được thu thập phải luôn cho phép tái tạo dữ liệu đối thủ trong tập dữ liệu huấn luyện, thử nghiệm hoặc xác thực tính hợp lệ. Ngoài ra, các phần thông tin phù hợp nhất có thể được sử dụng cho các kiểm chuẩn và chiến lược phòng thủ hơn nữa. Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

VÍ DỤ: Đối với dữ liệu đối thủ, thông tin cần được tài liệu hóa bao gồm, nhưng không giới hạn với, liên kết khu vực (nghĩa là tập con của tập dữ liệu), đặc tính của đối thủ, chiến lược tấn công và siêu tham số.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình bảo trì

5.3.5 Sự giảm nhẹ

5.3.5.1 RAR_MIT.1

Sau mỗi lần chạy kế hoạch thử nghiệm (xem 5.3.4 RAR_EVL.1), hãy phân tích xem có yêu cầu thêm các chiến lược phòng thủ hay không. Dựa trên kết quả phân tích tác động (xem 5.3.3 RAR_LIA.10) và đánh giá độ bền vững (xem 5.3.4 RAR_EVL.4), phải xác định xem có cần thiết phải cải thiện hơn nữa độ bền vững trước các đối thủ đã xác định và loại nhiễu loạn của đối phương hay không (xem 5.3.2 RAR_TMA.1 và RAR_TMA.2).

Lớp yêu cầu	Rất được khuyến khích

Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình bảo trì

5.3.5.2 RAR_MIT.2

Ánh xạ các loại chiến lược phòng thủ phù hợp vào các loại nhiễu loạn của đối thủ quan trọng hoặc có vấn đề (xem 5.3.RAR_TMA.1 và RAR_TMA.2). Cân nhắc sự thay đổi trong hành vi hoặc độ nhạy cảm của mô-đun AI đối với các loại đối thủ hoặc nhiễu loạn đối thủ khác nhau, nó rất có thể sẽ chứng minh sự cần phải phát triển một tổ hợp phòng thủ, tức là sự hợp thành của các chiến lược phòng thủ có khả năng khắc phục những thiếu sót do kết quả chạy thử nghiệm không đạt yêu cầu. Cần phải nhấn mạnh rằng một chiến lược phòng thủ đơn lẻ (và thường là chuyên biệt hóa) không thể được kỳ vọng một cách hợp lý để tính toán chống lại nhiều loại đối thủ, chứ chưa nói đến tất cả, các loại đối thủ hoặc nhiễu loạn của đối thủ. Do đó, các thành phần ứng cử viên phải được xác định, lựa chọn và cập nhật thường xuyên để tổ hợp phòng thủ có hiệu quả trong việc tính toán phạm vi rộng lớn các mối đe dọa an ninh. Thông thường, phải xác định xem việc huấn luyện lại mô-đun AI là cần thiết hay mong muốn.

VÍ DỤ: Trước khi điều chỉnh huấn luyện hoặc kiến trúc của mô-đun AI để tăng cường độ bền vững với các cuộc tấn công của đối thủ, sẽ có lợi nếu xem xét khả năng áp dụng các chiến lược phòng thủ dựa trên phát hiện, sửa đổi và bổ trợ.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình bảo đảm chất lượng, Quy trình bảo trì

5.3.5.3 RAR_MIT.3

Danh mục các chiến lược phòng thủ phù hợp (xem RAR_MIT.2) phải được cập nhật cơ bản thường xuyên hoặc sau những thay đổi cơ bản. Do tốc độ nghiên cứu nhanh chóng trong lĩnh vực này, danh mục phòng thủ đã định phải được cập nhật cơ bản thường xuyên. Khoảng thời gian cập nhật nào là hợp lý hoặc sự kiện nào sẽ kích hoạt cập nhật danh mục chiến lược phòng thủ phải được định trước bởi người dùng xác.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình phân tích hệ thống, Quy trình bảo trì

5.3.5.4 RAR_MIT.4

Lựa chọn và thực hiện danh mục phòng thủ (xem RAR_MIT.2). Phòng thủ thích hợp nên được

lựa chọn thực hiện một cách chiến lược và hoàn chỉnh. Việc lựa chọn phải được biện minh và tài liệu hóa.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình thực hiện

5.3.5.5 RAR_MIT.5

Đối với các phòng thủ nên được sử dụng kết hợp (được gọi là "tập hợp phòng thủ"), các lập luận và bằng chứng phải được cung cấp để đạt hiệu quả tập hợp. Cần phải thường xuyên xác minh rằng sự kết hợp của các chiến lược phòng thủ hiện vẫn duy trì một cách hiệu quả. Ngoài ra, sự kết hợp mới của các phương pháp phòng thủ phải được chứng minh là không chỉ tương thích, mà còn mạnh hơn khi kết hợp với nhau thay vì cô lập.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm, Quy trình xác minh

5.3.5.6 RAR_MIT.6

Trước khi triển khai, hãy thử nghiệm mô-đun AI với tổ hợp phòng thủ đã lựa chọn như được trình bày chi tiết trong kế hoạch thử nghiệm (xem 5.3.4 RAR_EVL.1) và báo cáo các thay đổi đối với tổ hợp phòng thủ trước đó. Một mô-đun AI phải được thử nghiệm dựa trên các tập dữ liệu thử nghiệm gốc và dựa trên tất cả, hoặc một tập hợp con có liên quan của các mẫu đối thủ đã lưu (xem 5.3.2 RAR_TMA.5). Hơn nữa, bắt buộc phải thực hiện lặp lại kế hoạch thử nghiệm trên các môi trường thử nghiệm sử dụng danh mục đo lường và tấn công của đối thủ (xem 5.3.2 RAR_TMA.5) cũng như các đo lường và số liệu hiệu năng chung. Ở giai đoạn này, cũng nên xem xét các khía cạnh chất lượng AI khác chẳng hạn như tính dễ hiểu (xem DIN SPEC 92001-1).

CHÚ THÍCH: Sửa đổi tiếp theo tiềm ẩn của tổ hợp phòng thủ sẽ kích hoạt chạy kế hoạch thử nghiệm khác, tức là một thích ứng phòng thủ cơ bản thuộc về tiêu chí đưa vào đã thiết lập của kế hoạch thử nghiệm (xem 5.3.4 RAR_EVL.1).

VÍ DỤ 1: Để có được thước đo thực tế về khả năng phòng thủ của mô-đun AI, phải đánh giá chiến lược phòng thủ chống lại các cuộc tấn công không xác định hoặc không nhìn thấy.

VÍ DỤ 2: Các phòng thủ phải được thử nghiệm chống lại các cuộc tấn công được thiết kế đặc biệt với trí tuệ phòng thủ đang tiến triển của mô-đun AI, được gọi là đối thủ thích ứng: vấn đề mà đối thủ có thể hiểu biết về phòng thủ được sử dụng và thích ứng với cuộc tấn công tương ứng của chúng phải được trang bị và đánh giá, với danh mục đo lường và tấn công của đối

thủ yêu cầu cập nhật liên tục.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác minh, Quy trình xác nhận tính hợp lệ

5.3.5.7 RAR_MIT.7

Hợp nhất các cân nhắc ở trên (xem RAR_MIT.1, RAR_MIT.2, RAR_MIT.3, RAR_MIT.4, RAR_MIT.5 và RAR_MIT.6) vào một danh mục tài liệu các thành phần của tổ hợp phòng thủ cùng với lý do tại sao các chiến lược phòng thủ cụ thể đã được lựa chọn và chứa tất cả thông tin liên quan về việc thực hiện và cập nhật kế hoạch thử nghiệm (xem 5.3.4 RAR_EVL.1). Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu (được gọi là "danh mục phòng thủ AR") cung cấp tổng quan và đặc điểm của các thành phần của tổ hợp phòng thủ cùng với lý do rõ ràng và ngắn gọn tại sao chúng được chọn để thực hiện. Đặc tính phòng thủ AR tóm tắt thêm tất cả các cập nhật cho kế hoạch thử nghiệm và về cơ bản bao gồm thông tin và thống kê tóm tắt liên quan đến thực hiện kế hoạch thử nghiệm.

CHÚ THÍCH: Các kết quả thu được sẽ cho phép so sánh: đó là, các kết quả phải cung cấp thông tin chi tiết rõ ràng về giá trị của mô hình như thế nào khi có và không có phòng thủ và giá phòng thủ khác nhau đối với tất cả các loại mẫu đối thủ có liên quan. Điều kiện tiên quyết để có thể so sánh là sử dụng nhất quán các đo lường và số liệu. Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

VÍ DỤ: Phòng thủ đang xem xét phải được tuyên bố rõ ràng, như là có các cuộc tấn công và các thiết lập tương ứng của chúng; phải nhấn mạnh vào vị trí truyền đạt rõ ràng của các số liệu được sử dụng để đánh giá.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình bảo trì

5.4 Yêu cầu cụ thể với sai lệch độ bền vững

5.4.1 Phạm vi, bối cảnh và tiêu chí

Các yêu cầu được liệt kê trong 5.2.2 hoàn toàn có thể áp dụng cho các cân nhắc về sai lệch độ bền vững.

5.4.2 Phân tích mô hình mối đe dọa

5.4.2.1 RCR_TMA.1

Xác định và cung cấp tính toán có cấu trúc về các loại có liên quan tiềm ẩn về sai lệch hoặc thay đổi phân phối đối với mô-đun AI. Đánh giá và tài liệu hóa những loại sai lệch hoặc thay đổi phân phối có thể đã gặp. Phát triển một cấu trúc phù hợp cho tài liệu của họ. Sự khác biệt

giữa môi trường huấn luyện và thử nghiệm cũng như giữa môi trường thử nghiệm và triển khai. Ngoài ra, sự bổ sung tài liệu với các nguồn tiềm tàng của những sai lệch hoặc thay đổi phân phối như vậy. Điều này đặt nền tảng công việc xác định và ngăn chặn hành vi không mong muốn hoặc thậm chí có hại gây ra bởi sai lệch hoặc thay đổi phân phối sau này.

VÍ DỤ 1: Các loại thay đổi phân phối có liên quan tiềm ẩn có thể được gọi là thay đổi đồng biến, thay đổi xác suất ưu tiên hoặc dịch chuyển khái niệm [12].

VÍ DỤ 2: Các nguồn chung của sự thay đổi phân phối là xu hướng lựa chọn mẫu và môi trường không cố định. Ví dụ, lĩnh vực lái xe tự động có rất nhiều vấn đề sai lệch tiềm ẩn: không chỉ các hiện tượng thời tiết như sương mù hoặc mưa ảnh hưởng đến chất lượng dữ liệu mà cả phần cứng cũng có thể làm cho xấu hơn do máy ảnh đưa ra sản phẩm không hoàn chỉnh.

VÍ DỤ 3: Trong lĩnh vực xử lý ngôn ngữ tự nhiên, các nguồn sai lệch chung nằm ở lỗi đánh máy hoặc ngôn ngữ mang phong cách riêng vì mô-đun AI thường được huấn luyện với đầu vào sạch hoặc được chuẩn hóa.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm, Quy trình phân tích hệ thống

5.4.2.2 RCR_TMA.2

Xác định thành công cho từng loại sai lệch hoặc thay đổi phân phối. Xây dựng trên định nghĩa về lỗi mô-đun AI (xem 5.2.2 RGE_SCC.1), cung cấp một định nghĩa phù hợp (có thể là một định nghĩa đang hoạt động) về sai lệch dữ liệu hoặc thay đổi phân phối thành công, tức là một định nghĩa gây ra cho mô-đun AI hoạt động sai. Phân tích phải được liên kết với các loại sai lệch hoặc thay đổi phân phối như được mô tả trong RCR_TMA.1.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm, Quy trình phân tích hệ thống

5.4.2.3 RCR_TMA.3

Xác định phạm vi của các chiến lược tăng cường dữ liệu. Chuẩn bị cho các chiến lược tăng dữ liệu có thể có khả năng tạo ra các loại hỏng dữ liệu sai lệch chung, tức là tạp nhiễu thành công hoặc là trên một phạm vi các kịch bản khác nhau hoặc là tất cả các điểm dữ liệu liên quan đến một kịch bản cụ thể (xem 5.2.2 RGE_SCC.1).

VÍ DỤ: Đặt tên rõ ràng các phân phối mà từ đó tạp nhiễu được lấy mẫu (ví dụ: Gaussian hoặc Student's t).

Lớp yêu cầu	Được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Nền tảng, Môi trường

Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình lập kế hoạch đề án, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng

5.4.2.4 RCR_TMA.4

Chỉ định cách phân biệt loại dữ liệu sai lệch hoặc thay đổi phân phối có thể được biểu thị hoặc mô phỏng trong các môi trường thử nghiệm khác nhau. Dựa trên đặc điểm của các loại dữ liệu sai lệch hoặc thay đổi phân phối trong RCR_TMA.1, hãy mô tả cách chúng có thể được thể hiện hoặc mô phỏng trong các môi trường thử nghiệm khác nhau, có thể là vật lý, kỹ thuật số, hoặc mô phỏng. Đối với từng môi trường thử nghiệm đang được xem xét, hãy sửa đổi và diễn giải lại các khái niệm về thành công và tầm quan trọng (xem RCR_TMA.2) của điểm dữ liệu bị sai lệch hoặc tạp nhiễu.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình phân tích hệ thống

5.4.2.5 RCR_TMA.5

Hợp nhất các cân nhắc ở trên (xem RCR_TMA.1, RCR_TMA.2, RCR_TMA.3 và RCR_TMA.4) vào danh mục các loại dữ liệu sai lệch hoặc thay đổi phân phối, chỉ số thành công và các ngưỡng, thừa nhận các kịch bản và môi trường thử nghiệm khác nhau. Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu (được gọi là "đặc tính CR") bao hàm một phân tích cấp cao về mô hình mối đe dọa, tức là đặc tả các loại sai lệch dữ liệu hoặc thay đổi phân phối có thể có khả năng xảy ra. Đặc tính CR cần luôn phân biệt giữa các kịch bản và môi trường thử nghiệm khác nhau (xem 5.2.2 RGE_SCC.4). Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.3 Phân tích Khả năng xảy ra & Tác động

5.4.3.1 RCR_LIA.1

Đánh giá các hậu quả tiềm ẩn của một sai lệch thành công hoặc sự thay đổi phân phối đáng kể. Một phân tích trong lĩnh vực này nên được cấu trúc xung quanh các kịch bản được coi là quan trọng như được trình bày trong 5.2.2 RGE_SCC.4 và các loại sai lệch hoặc thay đổi phân phối có thể xảy ra (xem 5.4.2 RCR_TMA.1). Kiểm soát các khía cạnh này (như được nêu trong đặc tính CR, xem 5.4.2 RCR_TMA.5), phân tích sẽ xác định loại hành vi không mong muốn hoặc thậm chí có hại mà mô-đun AI có thể biểu hiện (xem 5.2.2 RGE_SCC.3).

CHÚ THÍCH: Tương tự như AR, tài liệu về các hậu quả tiềm ẩn có thể được cấu trúc theo các dòng tác động đến tính toàn vẹn hoặc tính khả dụng của mô-đun AI.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Nền tảng, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm, Quy trình phân tích hệ thống, Quy trình xác minh

5.4.3.2 RCR_LIA.2

Đưa ra ước tính khả năng xảy ra cho từng loại sai lệch hoặc thay đổi phân phối. Đánh giá như thế nào ở đó mỗi loại sai lệch hoặc thay đổi phân phối sẽ xảy ra trong bối cảnh các kịch bản được xác định là có liên quan. Tuy nhiên, với các kịch bản khác nhau đã cho (xem 5.2.2 RGE_SCC.4), đánh giá có thể đưa ra các ước tính khả năng xảy ra khác nhau đối với từng loại sai lệch hoặc thay đổi phân phối.

Lớp yêu cầu	Được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm, Quy trình phân tích hệ thống

5.4.3.3 RCR_LIA.3

Thực hiện phân loại rủi ro và các kịch bản ưu tiên, trong đó CR có liên quan đến các loại sai lệch hoặc thay đổi phân phối được xác định là rất quan trọng. Thực hiện phân loại rủi ro về các loại sai lệch được xác định trước của sự thay đổi phân phối dựa trên quy trình quản lý rủi ro như được mô tả trong 5.2 Yêu cầu và Hướng dẫn về Quản lý Rủi ro (cụ thể là 5.2.2 Phạm vi, Bối cảnh và Tiêu chí), với rủi ro được khái niệm hóa như là sản phẩm của thiệt hại gây ra bởi sai lệch trong chất lượng dữ liệu hoặc thay đổi phân phối (xem RCR-LIA.1) và khả năng xảy ra của chúng (xem RCR-LIA.2). Ánh xạ các loại dữ liệu sai lệch đã xác định hoặc thay đổi phân phối vào tập dữ liệu sẽ đưa ra dấu hiệu về các kịch bản, trong đó CR là rất cần thiết hoặc ít nhất là các kịch bản hầu như bị ảnh hưởng bởi các vấn đề sai lệch.

VÍ DỤ: Các loại sai lệch chẳng hạn như sương mù hoặc phản xạ ánh sáng trong giao thông đô thị cũng như khả năng xảy ra của chúng có thể được đặt trong mối quan hệ với tập dữ liệu sạch. Chẳng hạn, sự xuất hiện của sương mù chỉ hợp lý trong bối cảnh dữ liệu ngoài trời.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.3.4 RCR_LIA.4

Thiết lập một tài liệu chứa các nguồn của nhiều và đa dạng các loại sai lệch hoặc thay đổi phân phối, các thuật toán để tạo ra chúng, cũng như đo lường và số liệu chỉ ra hoặc định lượng độ bền vững về mặt này. Thiết lập một tài liệu (từ nay về sau được gọi là "danh mục sai lệch và thay đổi phân phối") bao gồm và liên quan đến các mục và khía cạnh sau:

- Các thuật toán để tạo dữ liệu sai lệch, tức là các kỹ thuật để tăng cường dữ liệu sạch hiện có bằng cách sử dụng phân phối xác suất hoặc (mẫu của) chính dữ liệu bị sai lệch;

- Các số liệu và đo lường có thể chỉ ra và định lượng, trong không gian đầu vào, loại và mức độ sai lệch dữ liệu hoặc thay đổi phân phối;

- Các số liệu và đo lường đưa ra dấu hiệu về lỗi của một mô-đun AI và định lượng độ lớn của lỗi này;

- Ngược lại, các số liệu và đo lường có thể đưa ra chỉ báo về CR và định lượng độ lớn mà đầu ra vẫn duy trì ổn định khi mô-đun AI phảiđối mặt với sai lệch dữ liệu hoặc thay đổi phân phối.

Danh mục phải được tối ưu hóa theo cách sao cho tất cả các kịch bản an toàn quan trọng và các môi trường thử nghiệm có liên quan đều được tính đến (xem 5.2.2 RGE-SCC.4). Mặc dù, lý tưởng, các mẫu huấn luyện dữ liệu sai lệch đã được thu thập và bao gồm trong danh mục, nên bổ sung cho phương pháp tiếp cận này bằng cách sử dụng phân phối xác suất để Được khuyến khích tạo ra một số lượng lớn hơn và đa dạng hơn các loại điểm dữ liệu bị sai lệch.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.3.5 RCR_LIA.5

Nếu dữ liệu tổng hợp là một phần của danh mục sai lệch và thay đổi phân phối, hãy lập luận và tài liệu hóa rõ ràng cách thức dữ liệu đó bắt chước thực tế cũng như lý do và mức độ nó mang lại lợi ích cho việc đánh giá độ bền vững. Nếu dựa vào phân phối nhiễu loạn tổng hợp hoặc được mô hình hóa, thì phải chỉ ra chúng giống với các hiện tượng hoặc hiệu ứng trong thế giới thực như thế nào, hoặc ít nhất, nếu chúng không thể được giả định là giống với các hiệu ứng xảy ra tự nhiên, thì chúng có thể được đưa vào sử dụng hợp lý như thế nào.

Lớp yêu cầu	Được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm, Quy trình xác nhận tính hợp lệ

5.4.3.6 RCR_LIA.6

Lựa chọn cẩn thận và thường xuyên cập nhật danh mục sai lệch và thay đổi phân phối (xem RCR_LIA.4). Để phù hợp với tốc độ nhanh chóng của tiến bộ khoa học và khám phá trong lĩnh vực độ bền vững, điều rất cần thiết là phải thường xuyên xem xét, tinh chỉnh và cập nhật

danh mục sai lệch và thay đổi phân phối. Một lịch trình cập nhật phù hợp cần được thiết lập, điều này cũng có thể dựa trên các sự kiện như đưa ra các mô-đun AI mới hoặc những thay đổi trong môi trường triển khai.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.3.7 RCR_LIA.7

Tiêu chí xác định và lập tài liệu cho CR đầy đủ, tức là xác định phạm vi có thể chấp nhận hợp lý cho các số liệu của danh mục sai lệch và thay đổi phân bổ (xem RCR_LIA.4). Phạm vi có thể chấp nhận hợp lý có thể được thu thập và xây dựng bằng cách xem xét các đo lường kiểm chuẩn cho các số liệu như được liệt kê trong tài liệu. Quá trình này cũng bao gồm việc xác định và tài liệu hóa phạm vi và khoảng thời gian có thể chấp nhận phù hợp.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quyết định, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.3.8 RCR_LIA.8

Hợp nhất các cân nhắc ở trên (xem RCR_LIA.1, RCR_LIA.2, RCR_LIA3, RCR_LIA4, RCR_LIA.5, RCR_LIA.6 và RCR_LIA.7) bằng cách mở rộng tài liệu được gọi là “đặc tính CR”. Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu được gọi là "đặc tính CR mở rộng (phân tích khả năng và tác động)"). Tài liệu này bao gồm phân tích các hậu quả tiềm ẩn, khả năng xảy ra lỗi mô-đun AI do thiếu độ bền vững đối với các loại sai lệch dữ liệu hoặc thay đổi phân phối đã biết, và phân loại rủi ro tương ứng. Đặc tính CR mở rộng (phân tích khả năng xảy ra và tác động) hơn nữa còn bao gồm danh mục các kịch bản an toàn quan trọng được ưu tiên (xem 5.2.2 RGE_SCC.4) cũng như danh mục sai lệch và thay đổi phân phối (xem RCR_LIA.4). Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

VÍ DỤ: Một kiểm chuẩn chung trong thị giác máy tính là hiệu năng của mô hình phân loại hình ảnh trên dữ liệu khi đối mặt với các sai lệch chung như tạp nhiễu Gaussian hoặc tạp nhiễu bắn ở các mức độ nghiêm trọng khác nhau.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.4 Đánh giá độ bền vững

5.4.4.1 RCR_EVL.1

Phát triển một kế hoạch thử nghiệm (CR) dựa trên đặc tính CR mở rộng (phân tích khả năng xảy ra và tác động) như được mô tả trong 5.4.3 RCR_LIA.8. Chuyển đổi các loại sai lệch và thay đổi phân phối đã được xác định như là nghiêm trọng và tùy thuộc vào phân loại rủi ro cũng như danh mục sai lệch và thay đổi phân phối được liên kết thành một kế hoạch thử nghiệm cụ thể có thể được thực hiện cho mô-đun AI nhất định. Ở đây, những cân nhắc sau đây phải được tính đến:

- Kế hoạch thử nghiệm phải chứa các tiêu chí đưa vào phù hợp, tức là các sự kiện kích hoạt áp dụng kế hoạch thử nghiệm phải được chỉ định.

- Kế hoạch thử nghiệm phải bao gồm tiêu chí để dừng sớm, nghĩa là các sự kiện hủy bỏ thử nghiệm phải được mô tả.

- Các kết quả đánh giá phải được tiêu chuẩn hóa, hoặc ít nhất là thực hiện so sánh được, để cho phép kết luận rõ ràng, với kế hoạch thử nghiệm hướng dẫn cam kết quan trọng với các kết quả thu được. Cụ thể, nó phải hỗ trợ trong việc đánh giá những gì được biết đến như là các trường hợp biên, tức là các loại sai lệch hoặc thay đổi phân phối có khả năng xảy ra thấp.

- Kế hoạch thử nghiệm phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu, Môi trường
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình lập kế hoạch đề án, Quy trình quản lý quyết định, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm

5.4.4.2 RCR_EVL.2

Tích hợp một quy trình cập nhật thường xuyên cho kế hoạch thử nghiệm (xem RCR_EVL.1), tức là thường xuyên tiến hành xem xét sự hiện đại nhất và quyết định xem các thay đổi cơ bản đã xảy ra hay chưa (các loại thay đổi có liên quan phải được người dùng xác định trước). Do tốc độ nghiên cứu trong lĩnh vực này diễn ra nhanh chóng, nên danh mục sai lệch và thay đổi phân phối phải được cập nhật thường xuyên. Độ bền vững sai lệch của mô-đun AI phải được đánh giá lại và thử nghiệm lại sau mỗi lần cập nhật danh mục sai lệch và thay đổi phân phối hoặc kế hoạch thử nghiệm.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống và phần mềm, Quy trình vận hành, Quy trình bảo trì

5.4.4.3 RCR_EVL.3

Sau mỗi lần chạy kế hoạch thử nghiệm (xem RCR_EVL.1), so sánh kết quả với phạm vi có thể chấp nhận được và kết quả kiểm chuẩn cho mọi loại sai lệch hoặc thay đổi phân phối (xem

5.4.3 RCR_LIA.7), cũng như kịch bản và môi trường thử nghiệm (xem 5.2 .2 RGE_SCC.4), và quyết định xem mô-đun AI có đủ bền hay không. Mọi vi phạm kiểm chuẩn đều phải dẫn đến việc xem xét lại chiến lược giảm thiểu và dừng sơ bộ, và thậm chí là xem xét lại việc triển khai mô-đun AI.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác minh

5.4.4.4 RCR_EVL.4

Hợp nhất các kết quả và thông tin ở trên (xem RCR_EVL.1, RCR_EVL.2 và RCR_EVL.3) dưới dạng một tài liệu có cấu trúc (được gọi là "đánh giá CR"), tức là thu thập và lưu trữ kỹ lưỡng các kết quả đánh giá và tất cả thông tin về các lần hoạt động và cập nhật kế hoạch thử nghiệm cũng như dữ liệu sai lệch đã tạo ra, để sau đó có thể sao chép. Các phần thông tin liên quan nhất về việc thực hiện và cập nhật kế hoạch thử nghiệm phải được thu thập và lưu trữ (ví dụ: thống kê tóm tắt). Ngoài ra, dữ liệu bắt nguồn từ, hoặc bị ảnh hưởng bởi sai lệch hoặc thay đổi phân phối phải được rút gọn thành một bản tóm tắt toàn diện về cả tạo ra và phân nhánh của chúng, với khả năng tái tạo cấu thành mục tiêu chính (ví dụ: đồ tạo tác đã lưu). Nói cách khác, thông tin được thu thập phải luôn cho phép sao chép, trong tập dữ liệu huấn luyện, thử nghiệm hoặc xác nhận tính hợp lệ. Ngoài ra, các phần thông tin phù hợp nhất có thể được sử dụng hơn nữa cho các kiểm chuẩn và chiến lược phòng thủ. Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình bảo trì

5.4.5 Sự giảm nhẹ

5.4.5.1 RCR_MIT.1

Dựa trên kết quả của kế hoạch thử nghiệm (xem 5.4.4 RCR_EVL.1), thiết lập một danh sách các chiến lược giảm thiểu tiềm năng để tính toán các loại sai lệch hoặc thay đổi phân phối cốt yếu (xem 5.4.2 RCR_TMA.1). Đối với mỗi sai lệch dữ liệu hoặc thay đổi phân phối đã được xác định như là có vấn đề trong khi chạy thử nghiệm, phát triển chiến lược giảm thiểu. Điều này bao gồm lý luận tại sao nó lại hiệu quả và cân nhắc có căn cứ cũng như nỗ lực thực tế.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình thực hiện, Quy trình phân tích hệ thống

5.4.5.2 RCR_MIT.2

Biên dịch một tập hợp giảm thiểu từ danh mục các chiến lược giảm thiểu tiềm năng (xem RCR_MIT.1). Xem xét sự thay đổi trong hành vi hoặc cảm nhận của một mô-đun AI đối với các loại sai lệch hoặc thay đổi phân phối khác nhau, rất có thể cần chứng minh sự cần thiết phải phát triển một tập hợp giảm thiểu, tức là sự hợp thành của các chiến lược giảm thiểu ở đó có khả năng tập hợp để khắc phục những thiếu sót bị phơi bày bởi các kết quả chạy thử nghiệm không đạt yêu cầu. Cần phải nhấn mạnh rằng một chiến lược giảm thiểu đơn lẻ (và thường là chuyên biệt) không thể được kỳ vọng một cách hợp lý để chống lại, chứ chưa nói đến tất cả, các loại sai lệch dữ liệu hoặc thay đổi phân phối. Do đó, các thành phần ứng viên phải được xác định, lựa chọn và cập nhật thường xuyên để tập hợp giảm thiểu có hiệu quả trong việc chống lại các mối đe dọa an toàn khác nhau.

VÍ DỤ: Một tập hợp giảm thiểu đầy hứa hẹn có thể hoạt động như được mô tả sau đây. Tùy thuộc vào mức độ nghiêm trọng của sai lệch hoặc thay đổi phân phối và mức độ liên quan đến an toàn của nhiệm vụ, hãy chọn một trong ba hành vi sau:

- Đăng nhập nội bộ mà không ảnh hưởng đến phân loại hoặc thực hiện hành động (ví dụ: không có phản hồi cho khách hàng của một cửa hàng trực tuyến nhưng có thông báo cho quản lý chất lượng);

- Đăng nhập nội bộ mà không ảnh hưởng đến phân loại, nhưng thông báo cho người dùng;

- Đăng nhập nội bộ, từ chối phân loại, tùy chọn thông báo cho người dùng.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý quyết định, Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác định các yêu cầu hệ thống/phần mềm, Quy trình xác minh

5.4.5.3 RCR_MIT.3

Tìm kiếm liên tục các chiến lược giảm thiểu mới và tốt hơn (đặc biệt nếu các loại sai lệch hoặc thay đổi mới được xác định trong khi triển khai). Rất khuyến khích theo dõi tiến trình trong các lĩnh vực nghiên cứu tương ứng để cải thiện các chiến lược giảm thiểu hiện tại hoặc thay thế chúng bằng các phòng vệ mới lạ có khả năng đối phó hiệu quả hơn với các loại sai lệch dữ liệu hoặc thay đổi phân phối mới.

VÍ DỤ: Một loại sai lệch mới được phát hiện có thể yêu cầu giảm thiểu phù hợp khác được bổ sung vào tập hợp giảm thiểu. Chẳng hạn, sự xuống cấp của phần cứng và dẫn đến sai lệch dữ liệu có thể được giải quyết bằng cách thêm các kỹ thuật khử tạp nhiễu dựa trên máy học vào mô-đun AI.

Lớp yêu cầu	Rất được khuyến khích
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Khái niệm, Phát triển, Triển khai, Vận hành
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình quản lý hệ thống, Quy trình phân tích, Quy trình bảo trì

5.4.5.4 RCR_MIT.4

Trước khi triển khai, hãy thử nghiệm mô-đun AI với các chiến lược giảm thiểu đã thực hiện đối với các loại sai lệch hoặc thay đổi có liên quan đã xác định bằng cách sử dụng kế hoạch thử nghiệm. Trước khi mô-đun AI được triển khai, hãy đánh giá các chiến lược giảm thiểu đã thực hiện bằng cách sử dụng kế hoạch thử nghiệm như được mô tả trong 5.4.4 RCR_EVL.1, với loại dữ liệu sai lệch tiềm ẩn mới lạ đối với mô hình tạo thành tiêu chí đưa vào của kế hoạch thử nghiệm. Luôn cố gắng làm cho các kết quả có thể tái sản xuất và so sánh được bằng cách tài liệu hóa quy trình một cách chính xác và rõ ràng.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình đo lường, Quy trình bảo đảm chất lượng, Quy trình xác minh, Quy trình xác định tính hợp lệ

5.4.5.5 RCR_MIT.5

Hợp nhất các cân nhắc ở ở trên (xem RCR_MIT.1, RCR_MIT.2, RCR_MIT.3 và RCR_MIT.4) vào một tài liệu liệt kê các thành phần của tập hợp giảm thiểu cùng với lý do tại sao các chiến lược giảm thiểu cụ thể đã được lựa chọn và hàm chứa tất cả thông tin liên quan đến việc thực hiện và cập nhật kế hoạch thử nghiệm (xem 5.4.4 RCR_EVL.1). Những cân nhắc ở trên phải được tóm tắt dưới dạng một tài liệu (được gọi là "danh sách phòng thủ CR") cung cấp tổng quan và đặc điểm của các thành phần của tập hợp phòng thủ cùng với lý do rõ ràng tại sao chúng được chọn để thực hiện. Đặc tính giảm thiểu CR tóm tắt hơn nữa tất cả các cập nhật với kế hoạch thử nghiệm và về cơ bản bao gồm thông tin và thống kê tóm tắt liên quan đến thực hiện kế hoạch thử nghiệm. Tài liệu phải được cập nhật theo một lịch trình cập nhật hợp lý và khả thi.

Lớp yêu cầu	Bắt buộc
Nhân tố/các nhân tố ảnh hưởng	Mô hình, Dữ liệu
Giai đoạn/các giai đoạn vòng đời	Phát triển
Quy trình/các quy trình vòng đời	Quy trình quản lý rủi ro, Quy trình quản lý thông tin, Quy trình bảo đảm chất lượng

6 Hướng dẫn thực hiện

Tiêu chuẩn này không tạo thành một chuỗi hành động kiểm soát thời điểm và cách thức các yêu cầu chất lượng AI riêng biệt được thực hiện mà đúng hơn bao gồm một tập hợp các yêu cầu về độ bền vững dành riêng cho AI. Tuy nhiên, các tổ chức chịu trách nhiệm phát triển và vận hành các mô-đun AI sẽ được hưởng lợi từ việc lập kế hoạch và thực hiện các bước đánh giá và cải tiến chất lượng AI một cách có hệ thống. Ví dụ: phân loại mô-đun AI được phát triển mới như là “rủi ro cao” hoặc “rủi ro thấp” (xem chương 4) và thực hiện các quy trình quản lý rủi ro chung phù hợp (ví dụ: dựa trên ISO 31000, xem điều 5.1), trước khi xác định kế hoạch thử nghiệm chi tiết đánh giá Độ bền vững trước đối thủ (AR) và Độ bền vững sai lệch (CR) của mô-đun AI mới, sẽ là thứ tự tự nhiên của các bước công việc. chất lượng AI đã định cần được xử lý như một quy trình liên tục thay vì thực hiện một lần theo kinh nghiệm (xem điều 5.2), các tổ chức nên lập kế hoạch công việc liên tục của họ liên quan đến đánh giá và cải tiến chất lượng AI tương ứng (ví dụ: bằng cách phát triển các biểu đồ quy trình đánh giá chất

lượng AI với lộ trình quyết định rõ ràng dựa trên các số liệu chính, chẳng hạn như báo cáo độ bền vững của AI cho các mô-đun AI riêng biệt).

Bên cạnh một kế hoạch có hệ thống về các bước công việc liên quan đến chất lượng AI đang diễn ra, các tổ chức thực hiện các yêu cầu về độ bền vững của AI sẽ được hưởng lợi từ việc xử lý có kỷ luật các tài liệu liên quan đến chất lượng AI. Ví dụ: Thông số kỹ thuật DIN này đề xuất cho AR (xem điều 5.2) tạo ra tài liệu “đặc tính AR”, "đặc tính AR mở rộng (phân tích khả năng xảy ra và tác động)", “danh mục đo lường và tấn công của đối thủ”, "đánh giá AR" và "Danh mục phòng thủ AR", đối với CR (xem điều 5.3) tạo ra tài liệu "đặc tính CR", "đặc tính CR mở rộng (phân tích khả năng xảy ra và tác động)", "danh mục sai lệch và thay đổi phân phối", "đánh giá CR" và " danh mục phòng thủ CR", và đối với đánh giá độ bền vững theo kinh nghiệm tổng thể (AR+CR) của một mô-đun AI cụ thể, một "kế hoạch thử nghiệm" riêng biệt. Ngoài ra, một khi kế hoạch thử nghiệm AI đã được triển khai trên thực tế, kết quả đánh giá độ bền vững theo kinh nghiệm phải được lưu lại một cách thích hợp, ví dụ như ở dạng “báo cáo độ bền vững của AI” hoặc “hồ sơ độ bền vững của AI” (được liên kết với một phiên bản cụ thể của một mô-đun AI riêng biệt). Việc tạo ra có kỷ luật và cập nhật liên tục các tài liệu như vậy nên có độ ưu tiên cao đối với các tổ chức, vì trong tương lai, các tài liệu như vậy có thể được yêu cầu trong nhiều bối cảnh khác nhau, chẳng hạn như đạt được duyệt theo quy định đối với mô-đun AI mới, tăng độ tin cậy của cả khách hàng và nhà phân tích thị trường chứng khoán, và có thể ghi lại rằng các bước thích hợp để đánh giá độ bền vững của AI đã được thực hiện trong các trường hợp AI thất bại dẫn đến thủ tục tố tụng tại tòa án về trách nhiệm pháp lý của AI.

Tham chiếu đến các tài liệu về độ bền vững của AI tập trung được đề cập trong Thông số kỹ thuật DIN này:

Đối với AR, Thông số kỹ thuật DIN này đề xuất tạo ra các tài liệu sau:

- Đặc tính AR (xem 5.3.2 RAR_TMA.6)
- Danh mục đo lường và tấn công của đối thủ (xem 5.3.3 RAR_LIA.5)
- Đặc tính AR mở rộng (phân tích khả năng xảy ra và tác động) (xem 5.3.3 RAR_LIA.10)
- Kế hoạch thử nghiệm (AR) (xem 5.3.4 RAR_EVL.1)
- Đánh giá AR (xem 5.3.4 RAR_EVL.4)
- Danh mục phòng thủ AR (xem 5.3.5 RAR_MIT.7)

Đối với CR, Thông số kỹ thuật DIN này đề xuất tạo ra các tài liệu sau:

- Đặc tính CR (xem 5.4.2 RCR_TMA.5)
- Danh mục sai lệch và thay đổi phân phối (xem 5.4.3 RCR_LIA.4)
- Đặc tính CR mở rộng (phân tích khả năng và tác động) (xem 5.4.3 RCR_LIA.8)
- Kế hoạch thử nghiệm (CR) (xem 5.4.4 RCR_EVL.1)
- Đánh giá CR (xem 5.4.4 RCR_EVL.3)
- Danh mục phòng thủ CR (xem 5.4.5 RCR_MIT.5)

THƯ MỤC TÀI LIỆU THAM KHẢO

[1] ISO/ IEC/IEEE 12207:2017. Systems and software engineering – Software life cycle processes. Tech.

rep. ISO, IEC and IEEE, 2017.

[2] ISO/IEC 2382:2015. Information technology – Vocabulary. Tech. rep. ISO and IEC, 2015.

[3] E. Tabassi et al. A Taxonomy and Terminology of Adversarial Machine Learning, 2019.

[4] I. Goodfellow et al. Attacking machine learning with adversarial examples. OpenAI Blog, 2017.

[5] S. Huang et al. Adversarial Attacks on Neural Network Policies, 2017.

[6] X. Yuan et al. Adversarial Examples: Attacks and Defenses for Deep Learning, 2018.

[7] J. Metzen et al. Universal Adversarial Perturbations Against Semantic Image Segmentation, 2017.

[8] A. Kurakin et al. Adversarial Attacks and Defences Competition, 2018.

[9] Requirements for machine learning-based Quality of Service Assurance for the IMT-2020 Network,

2018.

[10] D. Hendrycks and T. Dietterich Benchmarking neural network robustness to common corruptions

and perturbations, 2019.

[11] J. Quiñonero-Candela. et al. Dataset Shift in Machine Learning. The MIT Press, 2008.

[12] J. Moreno-Torres et al. A unifying view on dataset shift in classification. Pattern Recognition 45, 2012.

[13] N. Ford et al. Adversarial examples are a natural consequence of test error in noise, 2019.

[14] G. Ditzler and R. Polikar. Semi-supervised Learning in Nonstationary Environments, 2011.

[15] IEC Whitepaper Artificial Intelligence across Industries, 2019.

[16] L. Engstrom et al. Exploring the landscape of spatial robustness, 2019.

[17] Y. Chung et al. Unknown Examples & Machine Learning Model Generalization, 2018.

[18] ISO 31000:2018. Risk management – Guidelines. Tech. rep. ISO, 2018.

[19] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2014.

[20] A. Madry et al. Towards deep learning models resistant to adversarial attacks, 2017