BỘ THÔNG TIN VÀ TRUYỀN THÔNG

**VIỆN CÔNG NGHIỆP PHẦN MỀM VÀ NỘI DUNG SỐ VIỆT NAM**

--------------------------------

**THUYẾT MINH**

**Dự thảo Tiêu chuẩn quốc gia**

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

**Hà Nội – 10/2022**

**MỤC LỤC**

[1 Tên gọi và ký hiệu của dự thảo TCVN 3](#_Toc114758780)

[1.1 Tên tiếng Việt của dự thảo TCVN: 3](#_Toc114758781)

[1.2 Tên tiếng Anh của dự thảo TCVN: 3](#_Toc114758782)

[2 Đặt vấn đề 3](#_Toc114758783)

[2.1 Sự cần thiết xây dựng tiêu chuẩn về tính đáng tin cậy trong trí tuệ nhân tạo 3](#_Toc114758784)

[2.2 Tình hình hoạt động chuẩn hóa liên quan đến tính đáng tin cậy trong AI 5](#_Toc114758785)

[3 Sở cứ xây dựng các yêu cầu kỹ thuật 17](#_Toc114758786)

[3.1 Sở cứ chung để đưa ra các yêu cầu 17](#_Toc114758787)

[3.2 Các thuộc tính của tính đáng tin cậy của AI 20](#_Toc114758788)

[3.3 Các yêu cầu, chỉ dẫn thực thi tính đáng tin cậy trong AI 23](#_Toc114758789)

[3.4 Lựa chọn tài liệu, tiêu chuẩn tham chiếu để xây dựng các nội dung chính của dự thảo tiêu chuẩn về tính đáng tin cậy trong AI 27](#_Toc114758790)

[4 Giải thích nội dung dự thảo tiêu chuẩn 30](#_Toc114758791)

[4.1 Xây dựng dự thảo tiêu chuẩn 30](#_Toc114758792)

[4.2 Các nội dung chính của dự thảo tiêu chuẩn 31](#_Toc114758793)

[5 Khuyến nghị áp dụng tiêu chuẩn 37](#_Toc114758794)

[6 Kiến nghị về hướng phát triển tiếp theo 39](#_Toc114758795)

**DANH MCỤ BẢNG BIỂU**

[Bảng 1 – Bảng đối chiếu nội dung dự thảo tiêu chuẩn với tài liệu tiêu chuẩn tham chiếu ISO/IEC 24028:2020 34](#_Toc115102173)

# Tên gọi và ký hiệu của dự thảo TCVN

## Tên tiếng Việt của dự thảo TCVN

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

## Tên tiếng Anh của dự thảo TCVN

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

# Đặt vấn đề

## Sự cần thiết xây dựng tiêu chuẩn về tính đáng tin cậy trong trí tuệ nhân tạo

Thế giới ngày nay đang chứng kiến ​​một cuộc cách mạng mới dựa trên một loạt các công nghệ trụ cột tiên tiến. Trong đó công nghệ Trí tuệ nhân tạo (AI) được coi là có tiềm năng thay đổi hầu hết mọi lĩnh vực như trong các ngành công nghiệp, sản xuất, kinh doanh, giáo dục, y tế, hoạt động của các bộ máy của chính phủ v.v.. Khả năng quan trọng tách biệt AI khỏi nhiều đột phá khoa học hoặc kỹ thuật trong quá khứ về giá trị mà nó có thể tạo ra là khả năng học máy. Máy móc có khả năng xử lý, theo dõi và trích xuất thông tin chi tiết, có giá trị cho nhiều mục đích khác nhau từ hàng triệu tập dữ liệu một cách rất nhanh chóng. Trí tuệ nhân tạo thực chất là một tập hợp các công nghệ và kỹ thuật liên quan như học máy, học sâu, xử lý ngôn ngữ tự nhiên, mang nơ-ron v.v.. có thể được sử dụng để bổ sung cho các phương pháp tiếp cận truyền thống về xử lý bởi các kỹ thuật khác nhau hoặc thực hiện bởi trí tuệ con người. AI là sự mô phỏng các quá trình thông minh của con người bằng máy móc, đặc biệt là hệ thống máy tính. Các quá trình này bao gồm học tập (thu nhận thông tin và các quy tắc được thể hiện dưới dạng thuật toán để sử dụng thông tin), suy luận (sử dụng các quy tắc để đạt được kết luận gần đúng hoặc rút ra kết luận), học tập thích ứng & các cơ chế tự điều chỉnh của nó.

AI là sự đổi mới quan trọng nhất đối với xã hội trong giai đoạn hiện nay. Nó có tiềm năng to lớn trong việc cải thiện chất lượng cuộc sống, có thể được sử dụng trong hầu hết các khía cạnh của cuộc sống của con người như dịch vụ chăm sóc sức khỏe, khu vực công, giáo dục, điện tử, ngân hàng v.v.. AI với đặc trưng là thuật toán ngày càng phức tạp hiện đang ảnh hưởng đến cuộc sống và nền văn minh của chúng ta hơn bao giờ hết. Ví dụ, hệ thống AI có thể giảm nhu cầu không mong muốn về tài nguyên bằng cách giám sát và quản lý chính xác dữ liệu về nhu cầu năng lượng của xã hội. Điều này sẽ dẫn đến việc xây dựng các cơ sở hạ tầng mang tính hiệu quả và tổ chức mạng lưới cung ứng một cách thông minh. Tương tự như vậy, việc kết hợp kiến thức của con người về điều trị, giải phẫu người, dược học với AI có thể cung cấp một cách tiếp cận mới để tận dụng kinh nghiệm, tri thức cho việc chuẩn đoán, tiên lượng trong hoạt động y tế, chăm sóc sức khỏe cũng như phát triển các loại thuốc mới làm giảm các sai sót, nâng cao chất lượng, rút ngắn thời gian và chi phí giảm một cách đáng kể của các hoạt động nói trên. Công nghiệp 4.0 áp dụng AI đối với các máy móc và quy trình được kết nối có khả năng giải quyết các tình huống phức tạp và có thể được tích hợp với các quy trình tự động hóa thông minh để phát triển các giải pháp tiên tiến trong các lĩnh vực. Đóng góp lớn nhất của AI sẽ là đối mặt và giải quyết các thách thức toàn cầu được nêu trong Mục tiêu Phát triển Bền vững của Liên hợp quốc (SDG), một tập hợp gồm 17 mục tiêu toàn cầu được thiết kế để đạt được tương lai bền vững cho tất cả mọi người. Để đạt được những mục tiêu này, sự đổi mới trong hệ thống AI hiện tại là điều tối quan trọng để chúng bao gồm quan điểm và chức năng nhân đạo trong xã hội nhằm hỗ trợ và mở rộng phúc lợi con người.

Trí tuệ nhân tạo thường được coi là một công nghệ hộp đen với một lỗi e ngại rằng liệu nó có được sử dụng để thao túng chúng ta hay không. Hiện nay câu hỏi đặt ra là làm thế thế nào để có thể vượt qua những thách thức này? Câu trả lời là cần phải xây dựng được các hệ thống, sản phẩm, dịch vụ AI có tính đáng tin cậy. AI đáng tin cậy có nghĩa là AI hợp pháp, đạo đức, kỹ thuật mạnh mẽ và tin cậy. Việc tạo ra trí thông minh nhân tạo đáng tin cậy và có đạo đức đòi hỏi sự hiểu biết không chỉ về bản thân công nghệ mà còn cả các điều kiện xã hội và đạo đức hiện tại, cũng như cách tính toán và đánh giá một cách thích hợp tác động của chúng đối với cách AI được thiết kế, xây dựng và thử nghiệm cũng như cách thức con người tương tác với nó. Để có sự tin tưởng hoàn toàn giữa công đồng xã hội và các hệ thống AI thì cả kiến ​​trúc thiết kế bên trong của AI, các ứng dụng cũng như giao diện con người sử dụng AI cần phải được xác định rõ ràng theo các nguyên tắc về độ tin cậy.

Trong bối cảnh của một hệ thống công nghiệp hoặc những thành phần được sử dụng cho một hệ thống công nghiệp, tính đáng tin cậy có nghĩa là một chủ thể đáng được tin cậy hoặc có thể được tin cậy. Sự tin cậy là tin tưởng vào điều gì đó hoặc tin tưởng vào ai đó, nhưng khái niệm này thường thiếu những hướng dẫn kỹ thuật, vì vậy Hiệp hội Internet công nghiệp (IIC) đã cải tiến định nghĩa để áp dụng khái niệm sự tin cậy và tính đáng tin cậy trong bối cảnh triển khai và vận hành các hệ thống công nghiệp. Để phục vụ mục đích sử dụng của các hệ thống, Kiến trúc tham chiếu Internet công nghiệp (IIRA) đã chỉ định năm đặc điểm chính liên quan sự tin cậy của một hệ thống nhằm đảm bảo rằng hệ thống có các chức năng hoạt động đầy đủ mà không bị ảnh hưởng bởi các yếu tố bên ngoài giới hạn. Năm đặc điểm này là: an toàn, bảo mật, độ tin cậy, khả năng phục hồi và quyền riêng tư. Vào đầu năm 2018, IIC đã điều chỉnh thuật ngữ này phù hợp với danh sách năm đặc điểm đề cập ở trên để xác định bản chất cốt lõi của tính đáng tin cậy. Đồng thời, Viện Tiêu chuẩn và công nghệ quốc gia của Hoa Kỳ (NIST) cũng nêu khái niệm này với các đặc điểm tương tự. Ngoài các đặc điểm về thể hiện tính tin cậy ở trên, IIC cũng chỉ rõ bốn nhóm về mối đe dọa gây nguy hiểm cho một hệ thống đáng tin cậy. Từ cách tiếp cận được mô tả trên. Tính đáng tin cậy được định nghĩa như sau:

*Tính tin cậy là mức độ tin tưởng một hệ thống thực hiện như mong đợi. Các đặc điểm của nó bao gồm tính an toàn, tính bảo mật, tính riêng tư, độ tin cậy và khả năng phục hồi khi hệ thống phải đối mặt với những xáo trộn của môi trường, lỗi của con người, lỗi hệ thống và các cuộc tấn công.*

Năm đặc điểm nêu trên là sự thể hiện các thuộc tính của tính đáng tin cậy. Có bốn loại mối đe dọa đối với tính đáng tin cậy là: xáo trộn môi trường, tấn công, lỗi con người và lỗi hệ thống. Những mối đe dọa này có thể dẫn đến các nguy cơ có thể dẫn đến những tổn thất của một hệ thống, có nghĩa là xuất hiện hệ quả tiêu cực gây ra do một những sự kiện không mong muốn hoặc không có trong kế hoạch.

Do vậy mục tiêu xây dựng tiêu chuẩn nhằm đáp ứng mục tiêu chức năng, công dụng, chất lượng: xác định các công nghệ, đặc điểm kỹ thuật ở mức tiên tiến và phù hợp để áp dụng tại Việt Nam. Cụ thể ở đây là các vấn đề, yêu cầu về tính tin cậy đối với các đối tượng là hệ thống sản phẩm, dịch vụ AI – Một loại hình sản phẩm mới hiện đang được triển khai ở nhiều lĩnh vực của đời sống kinh tế, xã hội trong nước ở các yêu cầu mang tính chỉ dẫn cho các vấn đề thuộc phạm vi dưới đây:

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

## Tình hình hoạt động chuẩn hóa liên quan đến tính đáng tin cậy trong AI

### Hoạt động chuẩn hóa của Viện Kỹ thuật Điện và Điện tử (IEEE) của Mỹ

Những năm vừa qua IEEE đã triển khai nhiều dự án nhằm xây dựng và ban hành các tiêu chuẩn trong lĩnh vực AI, trong đó có các vấn đề liên quan đến các khía cạnh thực hiện tính đáng tin cậy của AI. Dưới đây là các thông tin, kết quả hoạt động chuẩn hóa của IEEE liên quan vấn đề chuẩn hóa tính đáng tin cậy của AI.

#### Dự án IEEE P7001 - Tính minh bạch của các hệ thống tự trị

Dự án này được xây dựng nhằm mục đích xây dựng tiêu chuẩn phát triển các công nghệ AI cho việc xây dựng hệ thống tự quản, để có thể đánh giá các hành động của chính các hệ thống đó. Nó cung cấp các tính năng giúp người dùng hiểu tại sao một công nghệ, thuật toán, mô hình đưa ra các quyết định nào đó trong các tình huống khác nhau. Dự án này cũng nhằm xây dựng tiêu chuẩn về cách thức cung cấp tính minh bạch và trách nhiệm giải trình cho một hệ thống để giúp trợ giúp cải thiện hệ thống, chẳng hạn như kết hợp một bộ ghi dữ liệu sự kiện trong ô tô tự lái hoặc truy cập dữ liệu từ các cảm biến của thiết bị.

#### Tiêu chuẩn 2801-2022 – IEEE

**Tên tiêu chuẩn:** Khuyến nghị thực hành quản lý chất lượng các bộ dữ liệu cho trí tuệ nhân tạo trong y tế (Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence)

**Tình trạng:** Tiêu chuẩn được ban hành chính thức

Mục đích của tiêu chuẩn là thiết lập các quy tắc quản lý chất lượng của bộ dữ liệu cho trí tuệ nhân tạo y tế và cải thiện chất lượng dữ liệu tổng thể. Nội dung tiêu chuẩn đưa ra các thực hành được khuyến nghị nhằm xác định các thực tiễn tốt nhất để thiết lập hệ thống quản lý chất lượng cho bộ dữ liệu được sử dụng cho các thiết bị y tế sử dụng trí tuệ nhân tạo. Tiêu chuẩn đề ra mục tiêu chất lượng cho các tổ chức chịu trách nhiệm về bộ dữ liệu. Nó xác định một chu trình quản lý tập dữ liệu đầy đủ, bao gồm các khoản mục nhưng không giới hạn ở việc thu thập, truyền, sử dụng, lưu trữ, bảo trì và cập nhật dữ liệu. Các thực hành được khuyến nghị bao gồm danh sách các yếu tố quan trọng ảnh hưởng đến chất lượng của bộ dữ liệu, chẳng hạn như không giới hạn về nguồn dữ liệu, chất lượng dữ liệu, chú thích, bảo vệ quyền riêng tư, trình độ nhân sự / đào tạo / đánh giá, công cụ, thiết bị, môi trường, quy trình kiểm soát quy trình và các tài liệu liên quan.

#### IEEE – P 7000-2021

**Tên dự án**: Quá trình mô hình tiêu chuẩn đề cập đến các quan tâm về đạo đức trong giai đoạn thiết kế hệ thống (Standard Model Process for Addressing Ethical Concerns during System Design)

**Tình trạng:** Tài liệu phục vụ tiêu chuẩn hóa thuộc dự án tiêu chuẩn IEEE 7000™ - 2021

Tài liệu đưa ra tập hợp các quá trình mà một tổ chức cần có để xem xét các giá trị đạo đức trong suốt các giai đoạn khám phá, phát triển khái niệm để thiết kế các hệ thống AI. Tài liệu hướng dẫn triển khai các hoạt động quản lý và kỹ thuật trong giao tiếp minh bạch với các bên liên quan được chọn trong quá trình thiết kế để khuyến khích, thúc đẩy và ưu tiên các giá trị đạo đức liên quan đến việc truy xuất nguồn gốc các giá trị đạo đức thông qua việc xác định các loại hình giá trị của nó như khái niệm về vận hành; vai trò, vị trí và giá trị của từng bên trong thiết kế hệ thống. Tài liệu cũng mô tả các quy trình cung cấp khả năng truy nguyên các giá trị đạo đức, các yêu cầu đạo đức và thiết kế dựa trên rủi ro đạo đức. Tất cả các quy mô và loại hình tổ chức sử dụng mô hình vòng đời cần phải phù hợp với tài liệu này.

*Phạm vi:*

Tài liệu đưa ra tập hợp các quy trình mà các kỹ sư và nhà công nghệ cần phải xem xét về các giá trị đạo đức trong các giai đoạn khám phá, phát triển các khái niệm; bao gồm việc khởi tạo, phân tích và thiết kế hệ thống. Tài liệu cung cấp cho các kỹ sư và nhà công nghệ một quy trình có thể thực hiện nhằm điều chỉnh các quy trình quản lý đổi mới, các phương pháp thiết kế hệ thống, phương pháp kỹ thuật phần mềm để giải quyết các mối quan tâm hoặc rủi ro về đạo đức trong quá trình thiết kế hệ thống. Tuy nhiên tài liệu này không đưa ra hướng dẫn cụ thể về việc thiết kế các thuật toán để áp dụng các giá trị đạo đức như sự công bằng và quyền riêng tư.

*Mục đích:*

Mục đích của tài liệu này là cho phép các tổ chức thiết kế hệ thống với sự cân nhắc rõ ràng về các giá trị đạo đức xã hội và giá trị của cá nhân; chẳng hạn như tính minh bạch, tính bền vững, quyền riêng tư, công bằng và trách nhiệm giải trình, cũng như các giá trị thường được xem xét trong kỹ thuật hệ thống, chẳng hạn như hiệu năng và sự hiệu quả trong thiết kế.

#### IEEE – P7001/D4 – 2021

**Tên dự án:** Tiêu chuẩn về tính minh bạch của các hệ thống tự trị (IEEE Approved Draft Standard for Transparency of Autonomous Systems)

**Tình trạng:** Tài liệu dự thảo tiêu chuẩn phiên bản 4 thuộc dự án tiêu chuẩn IEEE 7001™ - 2021

Dự thảo tiêu chuẩn mô tả các mức độ minh bạch có thể đo lường, kiểm tra để đánh giá các hệ thống tự trị một cách khách quan và xác định mức độ tuân thủ.

*Phạm vi:*

Tiêu chuẩn này có thể áp dụng rộng rãi cho tất cả các hệ thống tự trị, bao gồm cả hệ thống vật lý và phi vật lý. Các ví dụ bao gồm xe có hệ thống lái xe tự động hoặc người máy hỗ trợ sinh hoạt (chăm sóc), hệ thống chẩn đoán y tế (người tư vấn) hoặc chatbot. Dự thảo tiêu chuẩn này đặc biệt quan tâm đến các hệ thống tự trị có khả năng gây hại. Do đó các hệ thống mà tính an toàn, tin cậy được coi là quan trọng nằm trong phạm vi của tài liệu này. Dự thảo xem xét các hệ thống có khả năng trực tiếp gây ra tổn hại về thể chất, tâm lý, xã hội, kinh tế hoặc môi trường, danh tiếng cá nhân, tổ chức. Các tác hại gián tiếp không ảnh hưởng trực tiếp tới các cá nhân, tổ chức cụ thể, chẳng hạn như những người không được phép truy cập trái phép vào dữ liệu bí mật, hoặc "tội ác không có nạn nhân" nhưng có tác động đến xã hội hoặc môi trường. Các hệ thống tự trị thông minh sử dụng học máy, các tập dữ liệu sử dụng để đào tạo các hệ thống như vậy cũng nằm trong phạm vi của dự thảo tiêu chuẩn này khi xem xét tính minh bạch của toàn bộ hệ thống. Tài liệu cung cấp một khuôn khổ để giúp các nhà phát triển các hệ thống tự trị có thể xem xét, nếu cần có thể áp dụng để thiết kế các tính năng các hệ thống đó để làm cho chúng minh bạch hơn. Khuôn khổ đề ra trong tài liệu nêu các yêu cầu đối với các tính năng tính minh bạch mà chúng mang lại cho hệ thống và cách chúng sẽ được chứng minh để xác định sự phù hợp với dự thảo tiêu chuẩn. Dự thảo tiêu chuẩn không cung cấp cho nhà thiết kế cách thức thiết kế tính minh bạch trong hệ thống của họ. Thay vào đó nó xác định một tập hợp các mức độ minh bạch có thể kiểm tra được và một tập hợp các yêu cầu tiêu chuẩn phải được đáp ứng để phù hợp với từng mức độ đưa ra. Các hệ thống tự trị và các quy trình mà chúng được thiết kế, thẩm định và vận hành sẽ chỉ minh bạch nếu các yêu cầu về tính minh bạch được đựa vào thiết kế của chúng. Ngoài ra, cần có các phương pháp kiểm tra, đo lường và so sánh các mức độ minh bạch khác nhau trong các hệ thống khác nhau. Lưu ý rằng tính minh bạch của hệ thống-hệ thống (tính minh bạch của hệ thống này với hệ thống khác) nằm ngoài phạm vi của dự thảo tiêu chuẩn. Tính minh bạch về cách các hệ thống con trong một hệ thống tự quản tương tác cũng nằm trong phạm vi của dự thảo tiêu chuẩn này.

*Mục đích:*

Mục đích của dự thảo tiêu chuẩn đặt ra các mức độ minh bạch có thể đo lường và kiểm tra được cho các hệ thống tự trị. Nguyên tắc chung đằng sau dự thảo tiêu chuẩn này là luôn có thể hiểu được tại sao và cách hệ thống hoạt động theo cách mà nó đã thực hiện. Tính minh bạch là một trong tám nguyên tắc chung được nêu trong thiết kế phù hợp về mặt đạo đức của IEEE và được nêu ra là “Cơ sở của một quyết định trong hệ thống tự trị và thông minh cụ thể nào đó phải luôn luôn có thể khám phá được”. IEEE đã thành lập một nhóm làm việc để soạn thảo dự thảo tiêu chuẩn này để đáp ứng trực tiếp khuyến nghị trong phần nguyên tắc chung của thiết kế phù hợp với đạo đức của IEEE.

#### IEEE – 7007-202

**Tên dự án:** Tiêu chuẩn bản thể luận cho các hệ thống tự động hóa và người máy điều khiển bằng đạo đức (Ontological Standard for Ethically Driven Robotics and Automation Systems)

**Tình trạng:** Tài liệu phục vụ tiêu chuẩn hóa của dự án tiêu chuẩn IEEE 7007TM - 2021

Tài liệu này thiết lập tập hợp các bản thể luận với các mức độ trừu tượng khác nhau, bao gồm các khái niệm, định nghĩa, tiên đề và các trường hợp sử dụng hỗ trợ việc phát triển các phương pháp luận theo định hướng đạo đức cho hoạt động thiết kế người máy và hệ thống tự động hóa. Nó tập trung vào lĩnh vực người máy và tự động hóa mà không xem xét bất kỳ ứng dụng cụ thể nào và có thể được sử dụng theo nhiều cách thức khác nhau. Ví dụ là các hướng dẫn hoặc tài liệu tham khảo trong quá trình phát triển người máy và hệ thống tự động hóa; phép giao tiếp rõ ràng và chính xác giữa các thành viên từ các cộng đồng khác nhau bao gồm người máy, hệ thống tự động hóa, đạo đức và các lĩnh vực liên quan. Người sử dụng tài liệu này cần có kiến ​​thức tối thiểu về lôgic học hình thức để hiểu tiên đề được thể hiện trong định dạng trao đổi logic chung.

*Phạm vi:*

Tiêu chuẩn này thiết lập một tập hợp các bản thể luận với các mức độ trừu tượng khác nhau chứa các khái niệm, định nghĩa, tiên đề và các trường hợp sử dụng được cho là có liên quan và thích hợp để thiết lập các phương pháp luận theo hướng đạo đức cho việc thiết kế người máy và hệ thống tự động hóa (R&A).

*Mục đích:*

Mục đích của tiêu chuẩn là thiết lập một tập hợp các định nghĩa và mối quan hệ của chúng để cho phép phát triển R&A phù hợp với các giá trị được chia sẻ và các nguyên tắc đạo đức được quốc tế chấp nhận nhằm tạo điều kiện tin tưởng vào việc tạo ra và sử dụng R&A. Nó nhấn mạnh vào sự phù hợp giữa đạo đức và kỹ thuật để cho phép cộng đồng hiểu cách thiết kế và triển khai các hệ thống này mang tính thực dụng trong bối cảnh của một xã hội dựa trên các giá trị (đạo đức). Những định nghĩa trong tài liệu cho phép giao tiếp chính xác giữa các chuyên gia toàn cầu thuộc các lĩnh vực khác nhau bao gồm người máy, tự động hóa, trí tuệ nhân tạo và đạo đức. Việc sử dụng các bản thể luận để biểu diễn tri thức trong bất kỳ lĩnh vực nào phải đạt được những mục tiêu: a) định nghĩa của các khái niệm được trong từng lĩnh vực cụ thể được mô tả một cách chính thức bằng ngôn ngữ cụ thể; b) các công cụ phân tích các khái niệm và mối quan hệ của chúng để tìm kiếm sự không nhất quán, không đầy đủ và nội dung thừa; c) ngôn ngữ được sử dụng trong quá trình giao tiếp giữa các người máy từ các nhà sản xuất khác nhau. Người sử dụng tiêu chuẩn này có trách nhiệm xem xét và tham chiếu các tiêu chí đạo đức thích hợp trong quá trình thiết kế hệ thống. Hơn nữa, người sử dụng tài liệu nên tham khảo tất cả các luật và quy định hiện hành.

#### IEEE - P2937 / D6

**Tên dự án:** Tiêu chuẩn dự thảo được IEEE phê duyệt để đo điểm chuẩn hiệu suất cho hệ thống máy chủ AI (Approved Draft Standard for Performance Benchmarking for AI Server Systems).

**Tình trạng:** Báo cáo kỹ thuật phiên bản 6 của dự án tiêu chuẩn P2937 của IEEE.

Điện toán trí tuệ nhân tạo (AI) khác với điện toán thông thường về cách hình thành, vận hành và cách sử dụng thiết bị. Hệ thống máy chủ AI, bao gồm máy chủ AI, cụm và cơ sở hạ tầng máy tính hiệu năng cao (HPC) được thiết kế đặc biệt cho mục đích này. Hiệu năng của các cơ sở hạ tầng này quan trọng đối với người dùng không chỉ trên các mô hình chung mà còn các cơ sở hạ tầng cho các lĩnh vực ứng dụng cụ thể. Tài liệu cung cấp các phương pháp chính thức để đo điểm chuẩn hiệu suất cho các hệ thống máy chủ AI; bao gồm các phương pháp tiếp cận để kiểm tra, đo lường. Ngoài ra, tài liệu còn cung cấp các yêu cầu kỹ thuật về các công cụ để đo điểm chuẩn.

**Phạm vi:**

Tài liệu kỹ thuật đưa ra các phương pháp kiểm tra hiệu năng của hệ thống máy chủ trí tuệ nhân tạo bao gồm máy chủ AI, cụm máy chủ AI và cơ sở hạ tầng điện toán hiệu suất cao AI (HPC).

### Liên minh viễn thông quốc tế ITU-T

Liên minh viễn thông quốc tế ITU-T đã thành lập các nhóm nghiên cứu trọng điểm (Focus Group – FG) để xúc tiến các hoạt động chuẩn hóa về AI. Dưới đây là một số kết quả của các hoạt động chuẩn hóa này liên quan đến các khía cạnh thực thi tính đáng tin cậy của AI.

#### DEL01: AI4H

**Tên tài liệu:** Các mối quan tâm về đạo đức (Ethics considerations).

**Tình trạng:** Tài liệu sơ thảo của nhóm nghiên cứu trọng tâm của ITU-T về AI trong y tế (FG-AI4H-201).

Tài liệu này được xây dựng với mục đích mô tả các chủ đề cần được giải quyết trong khái niệm “AI đối với những mối quan tâm về đạo đức y tế” có thể được phổ biến làm cơ sở để hình thành các nội dung, yêu cầu của các khuyến nghị trong tương lai đối với nhóm nghiên cứu trọng điểm về AI cho y tế của ITU-T. Công nghệ số, học máy và trí tuệ nhân tạo hứa hẹn tạo nên cách mạng hóa trong lĩnh vực y học, sức khỏe cộng đồng nhưng cũng làm nảy sinh các mối quan tâm đặc biệt về đạo đức, luật pháp và xã hội. Những vấn đề nảy sinh có thể là quyền truy cập công bằng, quyền riêng tư; cách thức sử dụng và người dùng phù hợp; trách nhiệm pháp lý, sự sự thiên vị và tính toàn diện của nó. Những vấn đề này có bản chất xuyên quốc gia vì việc thu thập, chia sẻ và sử dụng dữ liệu được tạo ra và / hoặc sử dụng bởi các công nghệ này vượt ra ngoài biên giới quốc gia. Các công cụ, phương pháp và công nghệ được sử dụng trong dữ liệu lớn và AI được kỳ vọng cải thiện các dịch vụ và hệ thống y tế. Tuy nhiên, nhiều câu hỏi vẫn chưa được giải đáp liên quan đến sự phát triển và sử dụng các công nghệ này có đạo đức, bao gồm cả việc các quốc gia có thu nhập thấp và trung bình sẽ được hưởng lợi như thế nào từ sự phát triển của AI. Một số cơ quan chính phủ, tổ chức học thuật, tổ chức phi chính phủ và các ủy ban chuyên trách về luật pháp các quốc gia đã bắt đầu tiếp cận các vấn đề đạo đức và thách thức do công nghệ số gây ra. Nhưng hiện tại vẫn chưa có hướng dẫn cụ thể nào về các trường hợp sử dụng. Do đó phát triển các hướng dẫn đạo đức hài hòa cho việc thiết kế và triển khai AI trong y tế ở phạm vi toàn cầu là nhu cầu cấp thiết.

#### DEL04: AI4H

**Tên tài liệu:** Đặc tính vòng đời phần mềm AI (AI software life cycle specification).

Tình trạng: Tài liệu sơ thảo của nhóm nghiên cứu trọng tâm của ITU-T về AI trong y tế (FG-AI4H-204).

Mục tiêu tài liệu hướng tới việc xác định, chuẩn hóa các vấn đề về:

a) Nhận dạng tất cả các tiêu chuẩn và thực tiễn tốt nhất có liên quan đến vòng đời của phần mềm ứng dụng trí tuệ nhân tạo. Quy trình vòng đời phần mềm AI cũng cần phải được xác định tương tự như các quy trình vòng đời phần mềm khác.

b) Đưa ra các phân tích và đánh giá quan trọng đối với các tài liệu liên quan để xác định các vấn đề cần phải giải quyết, chuẩn hóa

c) Xác định các bước cụ thể trong vòng đời đặc trưng phần mềm AI dùng trong y tế, chẳng hạn như các quy trình đào tạo và kiểm tra dựa trên dữ liệu tiềm năng cần được chú giải.

d) Đặc điểm kỹ thuật của phần mềm AI dùng trong y tế và xác định các thực hành tốt nhất đối với các bước vòng đời khác nhau.

#### DEL07 (FG-AI4H-J-027)

**Tên tài liệu:** Các quan tâm về đánh giá AI cho y tế (AI for Health Evaluation Considerations).

**Tình trạng:** Tài liệu phục vụ tiêu chuẩn hóa của nhóm nghiên cứu trọng tâm của ITU-T về AI trong y tế (FG-AI4H-204).

Tài liệu giới thiệu các vấn đề cần quan tâm trong việc đánh giá nhằm đảm bảo sự tin cậy của AI trong y tế và chăm sóc sức khỏe. Nó bao hàm 5 tài liệu liên quan là: DEL07.1 mô tả quy trình đánh giá; DEL07.2 các bài kiểm tra kỹ thuật; DEL07.3 các chỉ số kiểm tra; DEL07.4 đánh giá lâm sàng; DEL07.5 nền tảng đánh giá. Tài liệu giới thiệu các đặc điểm của quá trình đánh giá và thẩm định AI y tế mới nhất; đưa ra khái niệm điểm chuẩn đối chiếu cho mô hình tiêu chuẩn, các quan tâm sơ bộ về hoạt động trong quá trình đánh giá. Tài liệu cũng nêu các yêu cầu đối với nền tảng điểm chuẩn đối chiếu và các thực hành tốt nhất để đánh giá mô hình AI y tế được chọn từ nhiều nguồn khác nhay.

#### FGAI4AD-02

**Tên tài liệu:** Giao thức dữ liệu an toàn cho lái xe tự động - Các quan tâm về đạo đức và pháp lý của việc giám sát liên tục (Automated driving safety data protocol – Ethical and legal considerations of continual monitoring)

**Tình trạng:** Tài liệu thông báo kỹ thuật phục vụ tiêu chuẩn hóa của nhóm nghiên cứu trọng tâm của ITU-T về AI trong trợ giúp và tự động lái xe (FG- FGAI4AD).

Báo cáo kỹ thuật đề cập đến các vấn đề cần quan tâm về đạo đức và pháp lý của AI trong việc trợ giúp hoặc lái xe tự động. Các cân nhắc về vấn đề đạo đức liên quan đến quá trình giám sát liên tục và đảm bảo an toàn trong vận hành lái xe tự động, các khía cạnh liên quan đến trách nhiệm pháp lý được đề cập đến về cách thức giải quyết từ quan điểm của Hoa Kỳ và EU. Tài liệu cũng nêu một cách tổng quan về các quy định và tiêu chuẩn liên quan đến các khía cạnh đảm bảo an toàn, tin cậy cho hoạt động vận hành, trợ giúp hoặc tự động lái xe có ứng dụng AI.

### Viện Tiêu chuẩn Viễn thông châu Âu ETSI

Hiện tại hoạt động chuẩn hóa liên quan đến tính tin cậy trong AI của ETSI tập trung vào vấn đề bảo mật. Chính vì lý do đó ETSI đã thành lập nhóm nghiên cứu đặc biệt về bảo mật trí tuệ nhân tạo (ISG SAI) với mục tiêu nghiên cứu, xây dựng và ban hành các tiêu chuẩn kỹ thuật chất lượng cao để giải quyết những thách thức này. Mục tiêu đặt ra đối với nhóm ISG SAI là đưa ra các chức năng, yêu cầu về bảo mật của AI thông qua việc ban hành các tiêu chuẩn kỹ thuật; qua đó duy trì và cải thiện tính bảo mật của các công nghệ AI mới. Điều này có nghĩa là thực hiện một trong những khía cạnh quan trọng để đảm bảo tính tin cậy trong AI.

ISG SAI tập trung vào giải quyết vấn đề tiêu chuẩn hóa ở ba khía cạnh dưới đây về trí tuệ nhân tạo:

* Bảo vệ AI khỏi sự tấn công: trong đó AI là một thành phần trong hệ thống cần được bảo vệ
* Giảm thiểu những tác động tiêu cực của AI. Nghĩa là ở những nơi mà AI trở thành “vấn đề” (chẳng hạn như AI được sử dụng để cải thiện và nâng cao chức năng, hiệu quả của tấn công vector thông thường hoặc tạo các vectơ tấn công mới).
* Sử dụng AI để tăng cường các biện pháp bảo mật: bảo vệ hệ thống chống lại cuộc tấn công, trong đó AI được sử dụng và được coi là một phần của “giải pháp” bảo mật để cải thiện và nâng cao các biện pháp để đối phó với các cuộc tấn công, hoặc khắc phục các điểm yếu trong giải pháp bảo mật.

Cho đến thời điểm hiện tại, nhóm SAI đã ban hành các tài liệu dưới đây, là kết quả hoạt động chuẩn hóa về bảo mật trong AI của nhóm này.

#### ETSI GR SAI 006 V1.1.1 (2022-03)

**Tên tài liệu:** Bảo mật Trí tuệ nhân tạo (SAI); Vai trò của phần cứng trong bảo mật AI (Securing ArtificiaI Intelligence (SAI); The role of hardware in security of AI).

**Tình trạng:** Báo cáo kỹ thuật của nhóm SAI.

Tài liệu này xác định vai trò của phần cứng, bao gồm cả chuyên dụng và thông dụng trong bảo mật AI. Tài liệu đề cập đến các biện giảm thiểu khả dụng trong các phần cứng để ngăn chặn các cuộc tấn công và đưa ra các yêu cầu chung về phần cứng để hỗ trợ SAI. Ngoài ra tài liệu này cũng cung cấp một bản tóm tắt kinh nghiệm của các chuyên gia trong giới học thuật và lĩnh vực công nghiệp về bảo mật phần cứng cho AI nhằm giải quyết các lỗ hổng và điểm yếu trong các phần cứng mà từ đó xuất hiện khả năng tăng cường cại loại hình tấn công vectơ đối với AI.

#### ETSI GR SAI 001 V1.1.1 (2022-01)

**Tên tài liệu:** Bảo mật Trí tuệ nhân tạo (SAI); Bản thể luận tấn công AI (Securing Artificial Intelligence (SAI); AI Threat Ontology).

**Tình trạng:** Báo cáo kỹ thuật của nhóm SAI.

Tài liệu này xác định mối đe dọa Trí tuệ nhân tạo (AI) là gì và xác định cách phân biệt các mối đe dọa đối với AI với bất kỳ mối đe dọa nào không liên quan tới AI. Mô hình về các mối đe dọa đối với AI được trình bày dưới dạng bản thể luận để đưa ra cái nhìn về mối quan hệ giữa các tác nhân đại diện cho mối đe dọa, thực thể, tài sản bị đe dọa v.v.. Bản thể luận được mô tả trong tài liệu này là sự mở rộng cơ sở phân loại các mối đe dọa, tác nhân đe dọa được mô tả trong ETSI TS 102 165-1 để từ đó đưa ra cách thức giải quyết vấn đề tổng thể về bảo mật cũng như các chiến lược giảm thiểu môi đe dọa trong AI.

Bản thể luận được mô tả trong tài liệu này áp dụng cho các đối tượng AI vừa là tác nhân đe dọa vừa là mục tiêu tấn công.

#### ETSI GR SAI 002 V1.1.1 (2021-08)

**Tên tài liệu:** Bảo mật Trí tuệ nhân tạo (SAI); Bảo mật chuỗi cung ứng dữ liệu (Securing Artificial Intelligence (SAI); Data Supply Chain Security).

**Tình trạng:** Báo cáo kỹ thuật của nhóm SAI

Tài liệu này tóm tắt các phương pháp được sử dụng để tạo nguồn dữ liệu cho huấn luyện AI cùng với việc xem xét các sáng kiến ​​hiện có để phát triển các giao thức chia sẻ dữ liệu. Đồng thời, tài liệu nãy cũng đưa ra các phân tích về lỗ hổng đối với các phương pháp theo các sáng kiến nói trên để xác định phạm vi các yêu cầu có thể đối với các tiêu chuẩn nhằm đảm bảo tính toàn vẹn và an toàn đối với dữ liệu, thông tin và phản hồi được chia sẻ trong AI.

Tài liệu này hiện tại tập trung vào bảo mật dữ liệu và chưa đề cập đến bảo mật các mô hình AI. Các kỹ thuật để đánh giá và hiểu chất lượng dữ liệu cho các mục đích thực thi về hiệu năng, tính minh bạch hoặc đạo đức cũng được áp dụng cho việc đảm bảo an ninh. Mục đích của các đối tượng tấn công có thể là phá vỡ hoặc làm suy giảm chức năng của mô hình. Việc áp dụng các biện pháp giảm thiểu trong khía cạnh bảo mật có thể sẽ cải thiện hiệu năng, tính minh bạch và ngược lại.

Tài liệu hiện không đề cập đến các hành vi trộm cắp dữ liệu, được coi i là một vấn đề an ninh mạng truyền thống. Thay vào đó nó hướng tới xác định các hoạt động thao tác trên dữ liệu và ảnh hưởng của việc đó đối với các hệ thống học máy của AI.

#### ETSI GR SAI 005 V1.1.1 (2021-03)

**Tên tài liệu:** Bảo mật Trí tuệ nhân tạo (SAI); Báo cáo chiến lược giảm thiểu (Securing Artificial Intelligence (SAI); Mitigation Strategy Report).

Tình trạng: Báo cáo kỹ thuật của nhóm SAI.

Tài liệu tóm tắt và phân tích phương pháp giảm thiểu các mối đe dọa hiện có và tiềm năng để chống lại các mối đe dọa đối với các hệ thống dựa trên AI. Nó cung cấp thông tin về các kỹ thuật giảm thiểu các mối đe dọa đối với các các hệ thống chấp thuận triển khai các ứng dụng AI trên đó. Tài liệu cung cấp các mô tả để làm rõ các phương pháp bảo mật hệ thống dựa trên AI hiện để từ đó có các biện pháp giảm thiểu các mối đe dọa bảo mật đã biết hoặc tiềm ẩn. Nó cũng đề cập đến khả năng bảo mật, thách thức và hạn chế khi áp dụng các biện pháp giảm thiểu trong các hệ thống dựa trên AI trong một số trường hợp sử dụng tiềm năng.

### ISO/IEC

Các hoạt động xúc tiến chuẩn hóa về AI được thực hiện chủ yếu bởi nhóm nghiên cứu SC42 (chuyên về chuẩn hóa trong lĩnh vực AI và dữ liệu lớn) trực thuộc Ban kỹ thuật phối hợp 1 giữa ISO và IEC. Dưới đây là kết quả hoạt động chuẩn hóa về khía cạnh thực thi tính đáng tin cậy của AI thực hiện bởi các tổ chức này.

#### ISO/IEC TR 24028:2020

**Tên tiêu chuẩn:** Công nghệ thông tin – Trí tuệ nhân tạo – Tổng quan về tính đáng tin cậy trong trí tuệ nhân tạo (Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence).

**Tình trạng:** Tiêu chuẩn đã được ban hành

Tiêu chuẩn đề cập đến các vấn đề liên quan đến tính tin cậy trong các hệ thống, sản phẩm, dịch vụ AI, bao gồm các khía cạnh sau:

* Các phương pháp tiếp cận để thiết lập sự tin tướng vào các hệ thống, sản phẩm, dịch vụ AI thông qua tính minh bạch, khả năng giải thích, khả năng kiểm soát v.v..;
* Những cạm bẫy kỹ thuật và các mối đe dọa và rủi ro liên quan điển hình đối với hệ thống AI cùng với các kỹ thuật và phương pháp giảm thiểu có thể có;
* Các phương pháp tiếp cận để đánh giá và đạt được tính khả dụng, khả năng phục hồi, độ tin cậy, độ chính xác, an toàn, bảo mật và quyền riêng tư của các hệ thống, sản phẩm, dịch vụ AI

Các đặc tính kỹ thuật theo các mức độ đáng tin cậy đối với các hệ thống, sản phẩm, dịch vụ AI nằm ngoài phạm vi của tài liệu này.

#### ISO/IEC DIS 25059

**Tên tiêu chuẩn:** Kỹ thuật phần mềm— Yêu cầu và đánh giá chất lượng hệ thống và phần mềm (SQuaRE) - Mô hình chất lượng cho hệ thống AI (Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems)

**Tình trạng:** Tiêu chuẩn đang được phát triển ở dạng dự thảo

Dự thảo tiêu chuẩn này đề cập đến các khái niệm liên quan đến chất lượng, độ tin cậy của đối với các mô hình (phần mềm) sử dụng trong AI. Bằng cách tiếp cận theo mô hình chất lượng, các vấn đề liên quan đến sự tin cậy của các mô hình này được đề cập ở các khía cạnh dưới đây:

* Tính điều kiển được;
* Tĩnh mạnh mẽ
* Tính minh bạch
* Có thể can thiệp được

Tài liệu này có thể áp dụng cho các nhà thiết kế, phát triển phần mềm AI tham khảo để tạo dựng các mô hình phần mềm AI cho các mục đích sử dụng khác nhau. Tuy nhiên các vấn đề mang tính tổng thể về độ tin cậy của hệ thống, sản phâme dịch vụ AI là nằm ngoài phạm vi của dự thảo tiêu chuẩn này.

#### ISO/IEC DIS 23894

**Tên tiêu chuẩn:** Công nghệ thông tin – Trí tuệ nhân tạo – Quản lý rủi ro (Information technology – Artificial intelligence – Risk management).

**Tình trạng:** Tiêu chuẩn đang được phát triển ở dạng dự thảo.

Dự thảo tiêu chuẩn đưa ra các định nghĩa, khái niệm về quản lý rủi ro trong AI, là một trong các yếu tố ảnh hưởng tới tính tin cậy của các hệ thống AI. Các yếu tố rủi ro trong các giai đoạn tạo lập, thiết kế, phát triển, vận hành hệ thống AI được mô tả trong tài liệu. Các quá trình rủi ro có thể được quản lý thông qua các hoạt động được xác định trước, chẳng hạn như ghi nhận, thông báo, chuẩn đoán, theo dõi, đánh giá trong toàn bộ các quá trình hoạt động của một hệ thống AI. Theo các cách thức đề ra trong tài liệu, các rủi ro có thể được lường trước và ngăn chặn hoặc giảm thiểu sự xuất hiện nhằm đảm bảo sự hoạt động tin cậy của các hệ thống AI.

#### ISO / IEC TR 24029-1: 2021

**Tên tiêu chuẩn:** Trí tuệ nhân tạo (AI) - Đánh giá mức độ mạnh mẽ của mạng nơ-ron - Phần 1: Tổng quan (Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview)

**Tình trạng:** Báo cáo kỹ thuật được ISO/IEC xuất bản và lưu hành chính thức

Hiện tại tồn tại nhiều phương pháp thẩm định các hệ thống không phải AI, do vậy không phải lúc nào chúng cũng có thể áp dụng trực tiếp cho các hệ thống AI nói chung và mạng nơ-ron nói riêng. Tài liệu này đưa ta các phương pháp thẩm định tính mạnh mẽ của các mạng nơ-ron theo ba cách thức tiếp cận là phương pháp thống kê, phương pháp chính quy và phương pháp thực nghiệm.

Tuy nhiên, cần lưu ý rằng việc mô tả tính mạnh mẽ của mạng nơ-ron là một lĩnh vực nghiên cứu mở, do vậy tài liệu cũng chỉ ra những hạn chế của các phương pháp này trong cả hoạt động kiểm tra và thẩm định các mạng nơ-ron.

Do vậy mục tiêu của báo cáo kỹ thuật này cung cấp thông tin cơ bản về các phương pháp hiện có để đánh giá mức độ mạnh mẽ của mạng nơ-ron.

#### ISO/IEC AWI TS 12791

**Tên tiêu chuẩn:** Công nghệ thông tin – Trí tuệ nhân tạo – Xử lý sự thiên vị không mong muốn trong các tác vụ học máy phân loại và hồi quy (Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks).

**Tình trạng:** Tiêu chuẩn đang được xây dựng ở dạng tài liệu phục vụ chuẩn hóa.

Tài liệu này cung cấp các kỹ thuật giảm thiểu để xử lý tính thiên vị không mong muốn trong suốt vòng đời của hệ thống AI. Tuy nhiên các kỹ thuật này chỉ giới hạn đối với các hệ thống học máy để thực hiện các tác vụ phân loại và hồi quy. Phạm vi của tài liệu này có thể cho mọi lĩnh vực hoặc tổ chức sử dụng các tác vụ học máy phân loại và hồi quy.

### Nhận xét về hiện trạng hoạt động chuẩn hóa thuật ngữ, khái niệm trong Ai của các tổ chức tiêu chuẩn lớn trên thế giới

Những thông tin được khảo sát ở trên về hiện trạng tiêu chuẩn hóa về tính đang tin cậy trong AI của các tổ chức tiêu chuẩn hóa trên thế giới (IEEE, ITU-T, ETSI, ISO/IEC) cho thấy:

* Hoạt động tiêu chuẩn hóa về tính đáng tin cậy trong AI đã được các tổ chức triển khai chính thức trong những năm gần đây. Tuy nhiên, tuy từng mục tiêu hoạt động chuẩn hóa trong lĩnh vực này mà các tổ chức có cách tiếp cận khác nhau
* Đối với các tổ chức tiêu chuẩn IEEE, ITU-T, ETSI không xúc tiến dự án xây dựng và ban hành tiêu chuẩn riêng về tính đáng tin cậy trong lĩnh vực AI. Các khía cạnh cần xem xét về tính đáng tin cậy trong AI liên quan đến lĩnh vực cụ thể được xác định và bao hàm trong các tiêu chuẩn, đặc tính kỹ thuật, thông báo kỹ thuật được ban hành. Chẳng hạn như việc xác định các yêu cầu, thuộc tính đáng tin cậy trong AI được hiểu trong các lĩnh vực cụ thể như an toàn, bảo mật thông tin đối với các tài liệu của ETSI; tính minh bạch, yêu cầu về chất lượng dữ liệu đối với AI cho lái xe tự động, hệ thống tự trị đối với tài liệu của IEEE; Các mối quan tâm về đạo đức đối với các tài liệu của ITU-T. Do vậy việc xem xét đưa ra các yêu cầu về tính đáng tin cậy trong AI chưa được đề cập một cách tổng quát, mang tính toàn diện ở tất cả các mặt thể hiện của nó.
* ISO/IEC là tổ chức tiêu chuẩn quốc tế có hoạt động xây dựng tiêu chuẩn riêng trong lĩnh vực AI thông qua hoạt động của nhóm nghiên cứu SC42. Nhóm nghiên cứu này đã thực hiện dự án để xúc tiến chuẩn hóa về tính tin cậy trong các hệ thống AI. Với sự tập trung làm việc mang tính phối hợp của nhiều chuyên gia ở mọi lĩnh vực, trong đó có cả các chuyên gia từ các tổ chức tiêu chuẩn IEEE, ITU-T, ETSI trong dự án nói trên. Kết quả họ đã đưa ra một tài liệu, trong đó các vấn đề về thực thi tính tin cậy được xem xét một cách toàn diện đối với các hệ thống AI và có thể áp dụng trong phần lớn các lĩnh vực.

# Sở cứ xây dựng các yêu cầu kỹ thuật

## Sở cứ chung để đưa ra các yêu cầu

Tính đáng tin cậy và niềm tin vào tính đáng tin cậy là một khía cạnh thiết yếu của các hệ thống AI. Những hệ thống thể hiện đầy đủ tính đang tin cậy có thể coi là tổ hợp của các hệ thống thể hiện các đặc tính nổi về tính đáng tin cậy thông qua các tương tác giữa các hệ thống con của chúng. Các hệ thống con này có thể bao gồm công nghệ thông tin (CNTT), thường tập trung vào dữ liệu và công nghệ vận hành sử dụng dữ liệu (ở đây là công nghệ AI), cảm biến và thiết bị truyền động để thay đổi môi trường vật lý. Hậu quả của hành động, quyết định không chính xác có thể dẫn đến thiệt hại về nhân mạng, ảnh hưởng đến môi trường, gián đoạn cơ sở hạ tầng quan trọng hoặc các hậu quả khác như tiết lộ dữ liệu nhạy cảm, phá hủy thiết bị, tổn thất kinh tế hoặc tổn hại danh tiếng. Các mối quan tâm khác về mức độ đáng tin cậy đối với doanh nghiệp bao gồm việc tuân thủ các quy định, phòng tránh rủi ro về trách nhiệm pháp lý và xem xét các lợi ích tiềm năng từ việc duy trì một danh tiếng đáng tin cậy của một tổ chức, doanh nghiệp.

Tính đáng tin cậy thể hiện bởi các đặc điểm của chúng là vấn đề đang hết sức được quan tâm hiện nay đối với các sản phẩm, dịch vụ, hệ thống AI ở mọi lĩnh vực ứng dụng của chúng. Các đặc điểm về độ tin cậy và mối quan hệ giữa chúng là quan trọng trong tất cả các hệ thống, nhưng các ưu tiên tương đối và bản chất của các mối quan hệ phụ thuộc vào bối cảnh của ngành dọc và hệ thống đang được xem xét. Như vậy tính đáng tin cậy trong AI sẽ tác động đến rất nhiều các đối tượng, bao gồm cả các đối tượng sử dụng và đối tượng tạo lập ra các hệ thống AI. Các đối tượng này bao gồm các bên liên quan như thiết kế, phát triển, tích hợp hệ thống, chủ sở hữu, điều hành, các tổ chức, doanh nghiệp, cơ quan quản lý v.v.. và bất cứ bên liên quan nào ở trong và ngoài phạm vi hoạt động của những đối tượng nói trên quan tâm đến bảo mật, an toàn, độ tin cậy, khả năng phục hồi và quyền riêng tư của hệ thống vật lý mạng dựa trên AI.

Những cân nhắc, mối quan tâm được coi là quan trọng đối với tính đáng tin cậy bao gồm:

*Hậu quả:* Các ảnh hưởng tiêu cực có thể tác động trực tiếp hoặc gián tiếp hoặc kết hợp cả hai hình thức trên đến các bên liên quan. Ví dụ những thay đổi về chiến lược hoặc cách thức vận hành có thể ảnh hưởng trực tiếp đến hoạt động của các hệ thống hiện có. Tương tự như vậy các thay đổi về chủng loại sản phẩm hoặc hình thức hỗ trợ ảnh hưởng trực tiếp đến khách hàng sử dụng các dịch vụ, sản phẩm AI. Chúng cũng có thể ảnh hưởng gián tiếp đến khách hàng thông qua các yếu tố như chất lượng sản phẩm, dịch vụ. Hậu quả của những thay đổi trên tác động đến sự thay đổi về tính đáng tin cậy của các sản phẩm, dịch vụ AI ở một số khía cạnh hoặc tính tổng thể của nó. Điều này tác động trực tiếp đến mục tiêu cuối cùng của nhà cung cấp hệ thống, sản phẩm, dịch vụ AI đó là sự thay đổi về lượng khách hàng, doanh số hoặc lợi nhuận v.v..

*Khắc phục tức thì:* Hiệu ứng do các tác động đến tính đáng tin cậy có thể là rõ ràng đối với tất cả các bên ngay lập tức, chẳng hạn như các cuộc tấn công, các quyết định được đưa ra không đúng có thể gây ra các thiệt hại cho cả nhà cung cấp và người sử dụng. Nhưng đôi khi hiệu ứng do các tác động có thể không nhìn thấy rõ ràng và hậu quả là tiềm ẩn về sau này, chẳng hạn như hành vi trộm cắp tài sản trí tuệ, mô hình, dữ liệu v.v.. Sự chậm trễ trong việc phản hồi một sự kiện nào đó liên quan đến tính đáng tin cậy có thể gây ra tác hại cho đến khi nó được nhận ra. Do đó các cơ chế nhận biết và phản ứng tức thì đối với các mối đe dọa, các tác động cần được nhận biết một cách nhanh chóng và thực thi các phản ứng tức thì để giảm thiệu các hậu quả tiêu cực cho các bên liên quan.

*Mức độ nghiêm trọng:* Một hệ thống AI thất bại về độ tin cậy có thể gây ra các hậu quả ở mức độ từ rất nhỏ cho đến nặng nề như tê liệt hệ thống. Mức độ ảnh hưởng của nó đến các bên liên quan, hoặc đến các quy trình, mô hình là khác nhau tùy thuộc vào bối cảnh cụ thể. Ví dụ, một sai sót về bảo mật dữ liệu đầu vào mà không được phát hiện có thể dẫn đến việc các mô hình đưa ra dự báo, các quyết định sai lệch làm thiệt hai đến lợi ích của người dùng dịch vụ AI. Hệ thống, thành phần phần cứng cống có chức năng dự phòng có thể bị hỏng hóc làm ngừng hoạt động cung cấp dịch vụ. Sự sai sót có thể ít nghiêm trọng hơn và thực sự có thể thiện độ tin cậy nếu được nắm bắt và khắc phục trước khi xảy ra bất kỳ tai nạn hoặc hỏng hóc nghiêm trọng nào. Các yếu tố liên quan đến ngữ cảnh như đề cập ở trên có thể được xem xét đề ra các biện pháp hợp lý nhằm cải thiện mức độ tin cậy tổng thể của hệ thống AI.

Một số điểm bổ sung về tính đáng tin cậy trong những ngữ cảnh đáng lưu ý là:

*Sự khác biệt của các nhà cung cấp:* Các công ty (hay nhà cung cấp) khác nhau có thể sử dụng lựa chọn các công nghệ, thuật toán, tiêu chuẩn và thông lệ riêng tùy thuộc vào mục tiêu kinh doanh của họ. Do đó hệ thống, sản phẩm được thiết kế, trển khai, vận hành sẽ khác nhau về tiêu chí đảm bảo tính đáng tin cậy. Ngay cả khi các tiêu chuẩn, quy trình tương tự được sử dụng, các thông số của quy trình có thể khác nhau giữa các hệ thống, sản phẩm. Cùng với một yêu cầu đầu vào, đầu ra, thì việc thực hiện bên trong của các hệ thống có thể khác nhau. Do đó việc đánh giá rủi ro cũng như các môi quan tâm về khía cạnh thực hiện hệ thống sẽ khác nhau. Ở khía cạnh này, việc đề ra một khuôn khổ chung (hay gọi là tiêu chuẩn) về thực thi tính đáng tin cậy là cần thiết.

*Môi trường văn hóa, xã hội, luật pháp:* Tập quán, văn hóa và luật pháp địa phương chắc chắn ảnh hưởng đến mức độ đáng tin cậy đối với nhà cung cấp và người tiêu dùng bằng cách đặt ra các yêu cầu tối thiểu khác nhau. Một ví dụ là các yêu cầu về quyền riêng tư, trong đó việc sử dụng dữ liệu ở một quốc gia có thể được chấp nhận trong khi những điều đó lại là bất hợp pháp ở một quốc gia khác.

*Các bên liên quan bị ảnh hưởng:* Bất kỳ yếu tố đáng tin cậy nào cũng có thể ảnh hưởng đến nhiều bên liên quan. Ví dụ chống tấn công phá hoại mô hình liên quan đến sự hoạt động, vận hành an toàn của nhà cung cấp dịch vụ, sản phẩm AI. Tính tin cậy của dự báo ảnh hưởng đến người dùng v.v..

Có hai nguyên tắc chủ yếu để thực thi yếu tố về tính đáng tin cậy của hệ thống, sản phẩm, dịch vụ AI là:

*Nguyên tắc 1:* Các đặc điểm, thuộc tính về độ tin cậy phải được xem xét một cách tổng thể.

*Nguyên tắc 2:* Nhận thức sâu sắc bối cảnh sử dụng của hệ thống để đưa ra các biện pháp thực thi nhằm hài hòa hóa các mục tiêu thiết kế độ tin cậy và khả năng thực thi, tính năng thực hiện của hệ thống, sản phẩm, dịch vụ AI.

## Các thuộc tính của tính đáng tin cậy của AI

Tính đáng tin cậy được thể hiện bằng việc xác nhận nhiều lần các hành vi, quyết định có được sự tin cậy sự tin cậy đạt đượctrong một mối quan hệ lâu dài giữa các đối tượng, chủ thể chịu tác động của hệ thống dựa vào AI. Cụ thể hơn nữa tính đáng tin cậy có được thông qua việc đạt được ở cấp độ nào đó thuộc về các đặc điểm của nó phù hợp với bối cảnh hoạt động và sử dụng, có bằng chứng khẳng định các mức độ đó đang thực sự đạt được.

### Tính an toàn

Tính an toàn: *Là tình trạng hoạt động của một hệ thông không gây ra những rủi ro không thể chấp nhận được về sức khỏe, tinh thần của con người, trực tiếp hoặc gián tiếp gây ra thiệt hại về tài sản hoặc môi trường. (ISO / IEC Hướng dẫn 51: 2014).* Tính an toàn nhằm đảm bảo rằng một hệ thống hoạt động mà không xuất hiện rủi ro không thể chấp nhận được về tổn hại đối với sức khỏe, tinh thần của con người; thiệt hại đối với tài sản hoặc môi trường. Như vậy đảm bảo sự an toàn của hệ thống sẽ là các hoạt động nhằm giảm thiểu những rủi ro đó. Điều này đòi hỏi phải thực hiện các phân tích rủi ro, xác định phạm vi hoạt động an toàn và thiết kế hệ thống hoạt động trong các điều kiện ràng buộc được xác định để vận hành an toàn. An toàn có tính đến các các rủi ro bảo mật có thể cần các biện pháp bổ sung, chẳng hạn như cần tạo lập một hệ thống giám sát tách biệt với hệ thống được kiểm soát.

Các đặc điểm của tính an toàn trong hệ thống AI tương tác với nhau một cách linh hoạt dẫn đến tăng cường hoặc hạn chế lẫn nhau. Thiết kế hệ thống AI cần xem xét sự tương tác của các đặc tính này và tìm các giải pháp hài hòa để có được sự cân bằng giữa chúng. Các minh chứng thể hiện các giải pháp đó cùng với các lựa chọn hợp lý có thể cung cấp sự đảm bảo rằng hệ thống được thiết kế là đáng tin cậy cho một mục đích sử dụng nào đó. Ví dụ như mối quan tâm sử dụng hệ thống AI trong một trong một nhà máy sản xuất khác hệ thống AI sử dụng trong bệnh viện. Do đó đối với mỗi một hệ thống thì nhà thiết kế phải hiểu biết rõ ràng các yêu cầu về tính đáng tin cậy của hệ thống đó trong các hoạt động thiết kế, lắp ráp, vận hành và bảo trì để ở từng khâu, công đoạn để đạt được tính đáng tin cậy tổng thể của hệ thống đó.

Tính an toàn liên quan đến việc bảo vệ con người, trong khi tính bảo mật và tính tin cậy bảo vệ hệ thống và các đầu ra của nó. Khả năng phục hồi mô tả khả năng của hệ thống hoạt động khi điều kiện bình thường hoặc điều kiện mà độ tin cậy được kiểm soát bị mất.

### Tính bảo mật

Tính bảo mật: *Thuộc tính hoặc tài sản được bảo vệ khỏi sự truy cập, sử dụng, thay đổi hoặc phá hủy ngoài ý muốn hoặc trái phép, đảm bảo tính khả dụng, toàn vẹn và bí mật của chúng* (từ vựng IIC). Tính bảo mật của một hệ thống là hệ thống đó có các chức năng bảo vệ cho hệ thống khỏi sự truy cập, thay đổi hoặc phá hủy ngoài ý muốn hoặc trái phép. Tính bảo mật đảm bảo tính bí mật, tính toàn vẹn và tính khả dụng của dữ liệu của hệ thống ở mọi trạng thái, chẳng hạn như trạng thái không sử dụng, trạng thái di chuyển hoặc trạng thái đang được sử dụng; bảo bệ các hệ thống điều khiển, các mô hình, thuật toán; ứng dụng và dịch vụ AI tránh khỏi việc sử dụng hoặc hành vi không phù hợp, không đúng với các quy định. Mức độ ưu tiên của các mối quan tâm về bảo mật phụ thuộc vào các bối cảnh cụ thể. Tính bảo mật có thể đạt được bằng các biện pháp bảo vệ các thiết bị điểm cuối, thiết bị mạng, danh tính máy, máy ảo, vùng chứa cũng như các ứng dụng và dữ liệu. Trong các hệ thống AI công nghiệp, dữ liệu điều khiển được sử dụng để thực hiện các hoạt động vật lý có khả năng gây thiệt hại vật chất nên có thể cần các biện pháp bảo vệ tăng cường.

### Tính tin cậy

Ttính tin cậy: *Là khả năng của một hệ thống hoặc các thành phần của nó thực hiện các chức năng theo thiết kế ở các điều kiện xác định trong một khoảng thời gian vận hành xác định (ISO / IEC 27040: 2015).* Tính tin cậy thể hiện khả năng của một hệ thống hoặc các thành phần của nó thực hiện các chức năng theo thiết kế ở các điều kiện xác định trong một khoảng thời gian vận hành xác định. Điều này bao gồm mọi cân nhắc về sự xuống cấp vật lý, các phiên bản phần mềm hết hạn và các trục trặc tiềm ẩn dẫn đến việc phải thực hiện các hoạt động bảo trì thường xuyên hoặc định kỳ, thay thế các thành phần hệ thống bị lão hóa hoặc cập nhật thay thế phần mềm, dữ liệu mới. Tính tin cậy cho phép hệ thống hoạt động không bị gián đoạn, một yếu tố cần thiết để đảm bảo rằng hệ thống sẽ đáp ứng các kỳ vọng về mục đích, chức năng hoạt động trong vòng đời hoạt động của nó.

Để đạt được tính đáng tin cậy của một hệ thống thì trước hết các vấn đề liên quan đến hệ thống đó cần phải được định nghĩa một cách rõ ràng và đầy đủ, vì các định nghĩa này xác định các yêu cầu cụ thể đối với hệ thống cần được xây dựng. Ví dụ như việc xác định các vấn đề quan tâm đối với tính đáng tin cậy của hệ thống camera nào đó khác với việc xem xét mức độ đáng tin cậy đối với toàn bộ nền tảng của hệ thống AI cung cấp các dịch vụ liên quan tới các camera đó. Các rủi ro liên quan đến camera có thể bao gồm mất quyền riêng tư hoặc bí mật thương mại, mất đi sự giám sát an toàn, bảo mật hoặc xuất hiện các lỗ hổng tạo cơ hội cho hoạt động tấn công bảo mật. Trong khi đó đó đối với nền tảng AI cung cấp dịch vụ camera đó thì tính đáng tin cậy của hệ thống camera chỉ là một phần trong tổng thể tính đáng tin cậy của hệ thống đó, chẳng hạn như sự tin cậy của các hệ thống, mô hình, thuật toán nhận dạng, các thành phần phần cứng, phần mềm khác của hệ thống.

### Khả năng phục hồi

Khả năng phục hồi: *Là khả năng của một hệ thống hoặc thành phần của nó có thể duy trì các cấp độ dịch vụ có thể chấp nhận được khi đối mặt với sự cố (Từ vựng IIC).* Khả năng phục hồi cho phép một hệ thống hoặc thành phần của nó có thể duy trì các cấp độ dịch vụ có thể chấp nhận được khi đối mặt với sự cố. Ngược lại với độ tin cậy, khả năng phục hồi giải quyết các trạng thái xuất hiện mà hệ thống không mong muốn và không ở trong kế hoạch vận hành của nó, ví dụ như lỗi của con người trong quá trình vận hành hoặc sự kiện môi trường (mất điện, động đất, v.v.). Mục đích chính của khả năng phục hồi là ngăn ngừa hoặc giảm thiểu các tác động mang tính nghiêm trọng gây ra do sự gián đoạn hoạt động của hệ thống do hư hỏng, lỗi, mất hoạt động hoặc tác động của nhiều nguyên nhân khác.

### Tính riêng tư

Tính riêng tư: *Quyền của các cá nhân kiểm soát hoặc tác động đến những thông tin liên quan đến họ có thể được thu thập và lưu trữ, tiết lộ bởi ai và cho những đối tượng nào (ISO / TS 17574: 2009 2 được tham chiếu trong từ vựng IIC).* Tính riêng tư là sự bảo vệ quyền của các cá nhân trong việc kiểm soát các thông tin nào liên quan đến họ có thể được ai thu thu thập và lưu trữ và những thông tin đó có thể được tiết lộ cho ai và trong bao lâu.

Những đặc tính này tương tác với nhau theo nhiều cách thức, ví dụ như:

* Tính tin cậy đề cập đến sự hoạt động chính xác của hệ thống trong trong các điều kiện cụ thể, trong khi khả năng phục hồi đề cập đến duy trì sự hoạt động của hệ thống trong các điều kiện không theo kế hoạch.
* Tính riêng tư chỉ bảo vệ dữ liệu liên quan đến con người nhưng không đề cập đến dữ liệu kinh doanh hoặc dữ liệu hoạt động, mà các dữ liệu này được hiểu trong khái niệm về bảo mật.

### Trách nhiệm giải trình

Trách nhiệm giải trình: *Nghĩa vụ của một cá nhân hoặc tổ chức phải giải trình về các hoạt động của họ để hoàn thành một nhiệm vụ theo vị trí, vai trò được giao, nhận trách nhiệm về các hoạt động, công việc liên quan đến nhiệm vụ đó và công bố các kết quả thực hiện nhiệm vụ một cách minh bạch. (ISO / TS 21089: 2018)*

Trách nhiệm giải trình cần được cung cấp cho các bên liên quan, dưới dạng bằng chứng cho thấy một hệ thống có thể tin cậy được.

### Sự đảm bảo

Sự bảo đảm: *Sự bảo đảm là cơ sở khẳng định để tin tưởng rằng một yêu cầu đã hoặc sẽ đạt được (ISO / IEC 15026-1: 2013 2, được tham chiếu trong từ vựng IIC).* Hiểu bối cảnh của các khái niệm liên quan đến các khía cạnh cần đảm bảo của hệ thống AI là điều cần thiết để hỗ trợ thực thi tính đáng tin cậy cũng như các giải pháp mang tính hài hòa với các vấn đề khác trong thực thi hệ thống, dịch vụ AI.

## Các yêu cầu, chỉ dẫn thực thi tính đáng tin cậy trong AI

Các thuộc tính thể hiện tính đáng tin cậy của một hệ thống AI như đề cập ở trên cần phải được thực hiện ở những cấp độ nào đó và phải thực hiện theo cách tiếp cận mang tính cân bằng các khía cạnh thực thi của tính đáng tin cậy. Các thuộc tính về tính đáng tin cậy thể hiện ở nhiều khía cạnh thực thi trong suốt vòng đời của hệ thống AI và được thực hiện, xem xét, đánh giá, thẩm định, chấp nhận bởi nhiều bên liên quan. Mức độ đáng tin cậy có thể thay đổi để đạt được sự cân bằng phù hợp với nhà cung cấp và bối cảnh hoạt động của các hệ thống, dịch vụ AI. Ví dụ việc thiết kế hệ thống ra quyết định cho các cơ cấu chấp hành cần phải có các mô hình có độ chính xác cao (các mô hình dựa trên kịch bản xây dựng trước) trong khi các mô hình dự báo có thể thực thực hiện trên cơ sở suy luận mang tính thống kê, xác suất. Bảo mật trong dữ liệu điều tra thị trường có mức ưu tiên thấp hơn so với bảo mật thông tin, dữ liệu cá nhân v.v..

Bối cảnh của mức độ đáng tin cậy rất quan trọng vì việc giám sát mọi sự kiện bất lợi và hiểu được hậu quả của nó là mối quan tâm không thể thiếu trong việc thiết kế và vận hành hệ thống đạt được mức độ tin cậy. Các ảnh hưởng gián tiếp có thể khó phát hiện và kiểm soát hơn, chẳng hạn như nếu chúng nằm trong chuỗi cung ứng cho các thành phần của hệ thống. Các thuộc tính của tính tin cậy phải được xem xét đồng thời trong bối cảnh hệ thống, mối quan tâm của các bên liên quan và các hậu quả tiềm ẩn. Một hệ thống phải được xem xét một cách một tổng thể, và điều này bao gồm việc xem xét các tác động gián tiếp của các yếu tố thiết kế hoặc các thay đổi và cách những tác động gián tiếp này có thể ảnh hưởng đến độ tin cậy.

Trên cơ sở đó, các yêu cầu mang tính chỉ dẫn thực thi tính đáng tin cậy của một hệ thống AI cần phải được đưa ra. Chúng được xét trong bối cảnh tổng thể đối với một hệ thống AI, trong suốt quá trình vòng đời của nó, chẳng hạn như nghiên cứu, thiết kế, triển khai, vận hành và ngừng làm việc.

### Áp dụng các khuôn khổ hiện hành để thực thi tính đáng tin cậy

Hệ thống AI thường là các hệ thống hiện có được tăng cường khả năng AI. Do đó tất cả các phương pháp hiện hành tiếp cận thực thi giải quyết các mối quan tâm về tính đáng tin cậy của các hệ thống hiện có (chẳng hạn như đối với tính đáng tin cậy của hệ thống phần mềm, phần cứng, các quy trình quản lý chất lượng sản phẩm) đã được ban hành có liên quan cần được xác định, nhận biết để tiếp tục áp dụng cho hệ thống. Chúng bao gồm các phương pháp tiếp cận về chất lượng (cả dữ liệu đo đạc và phương pháp đo lường), tính an toàn, nguy cơ tổn hại và các khuôn khổ quản lý rủi ro (chẳng hạn như các khuôn khổ hiện có về tính bảo mật và tính riêng tư).

Do đó, tiêu chuẩn cần phải có những chỉ dẫn kỹ thuật để thực hiện các công việc như:

* Nhận biết các lớp tin cậy: chẳng hạn như lớp tin cậy về vật lý, hạ tầng mạng, thiết bị, phần mềm đầu cuối, lớp tin cậy của các đối tượng trong xã hội như văn hóa, bản sắc, niềm tin v.v..
* Áp dụng các tiêu chuẩn chất lượng phần mềm và dữ liệu: dựa trên cơ sở nhận biết các lớp tin cậy để từ đó áp dụng các tiêu chuẩn đã được ban hành để có thể đánh giá được mức độ tin cậy cho từng thuộc tính cụ thể.
* Áp dụng chính sách, quy trình quản lý rủi ro: Quản lý rủi ro là một quá trình phòng ngừa nhằm đảm bảo rằng sản phẩm AI hoặc dịch vụ AI “theo thiết kế” có tính đáng tin cậy trong suốt vòng đời của nó. Các chính sách, quy trình quản lý rủi ro cụ thể được ban hành hoặc cần được cải tiến nâng cao để phù hợp với hệ thống AI cần được áp dụng.

Xem xét áp dụng các biện pháp thực thi đáng tin cậy bằng các biện pháp phần cứng: Chẳng hạn như triển khai trên cơ sở hạ tầng điện toán, tăng cường biện pháp phần cứng để giảm thiểu tấn công, mã hóa mô hình, dữ liệu v.v..

### Các bên liên quan tới tính đáng tin cậy

Các bên liên quan khác nhau có thể có quan điểm khác nhau về tầm quan trọng đối với các thuộc tính khác nhau về tính đáng tin cậy của AI. Nội dung này cần phải xác định được các vấn đề như:

* Các bên liên quan khác nhau có thể bị ảnh hưởng như thế nào bởi công nghệ AI được triển khai trong một sản phẩm hoặc dịch vụ;
* Tài sản (xét theo khái niệm trong lĩnh vực AI) được định giá bởi các bên liên quan khác nhau như thế nào, được sử dụng hoặc bị ảnh hưởng như thế nào bởi việc sử dụng AI trong một sản phẩm hoặc dịch vụ;
* Việc sử dụng AI trong một sản phẩm hoặc dịch vụ có liên quan như thế nào đến các giá trị do các bên liên quan khác nhau nắm giữ.

Trên cơ sở đó các chỉ dẫn kỹ thuật của tiêu chuẩn cần đưa ra các nội dung như:

* Liệt kê các loại hình bên liên quan
* Tài sản trong lĩnh vực AI
* Giá trị được hiểu như thế nào

### Nhận biết các mối quan tâm cấp cao

Mối quan tâm cấp cao được hiểu là việc thực thi tính tin cậy ở quy mô tổ chức của các bên liên quan về vấn đề này. Trong đó cần nhấn mạnh đến xác định vai trò, trách nhiệm và trách nhiệm giải trình giữa các bên liên quan. Hệ thống AI có thể tồn tại trong cả một chuỗi giá trị thương mại quốc tế phức tạp và trên các khuôn khổ xã hội xuyên quốc gia. Do đó các bên liên quan cần có chia sẻ sự hiểu biết về trách nhiệm mà họ thực hiện đối với các bên liên quan khác.

Tính an toàn là một khía cạnh quan trọng của tính đáng tin cậy. Do đó việc xem xét các khía cạnh về tính an toàn được ưu tiên cao. Thông thường rủi ro gây hại của hệ thống được nhận thức càng cao thì yêu cầu về tính đáng tin cậy càng cao. Hệ thống AI cũng như bất kỳ hệ thống nào khác, được kỳ vọng sẽ không gây ra bất kỳ tác hại không có chủ ý nào. Điều này không chỉ bao gồm tác hại hữu hình (ví dụ, đối với sức khỏe của sinh vật, tài sản và môi trường vật chất) mà còn cả tác hại vô hình (ví dụ đối với môi trường xã hội và văn hóa).

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

Tính dễ bị tổn thương được nhận định là điểm yếu trong các thành phần, chức năng hệ thống, sản phẩm, dịch vụ dựa trên AI có thể có thể bị lợi dụng bởi một hoặc nhiều mối đe dọa. Chúng bao gồm sự phụ thuộc vào dữ liệu, tính không rõ ràng, không thể đoán trước của các mô hình học máy. Việc sử dụng dữ liệu có thể dẫn đến các mối đe dọa mới về bảo mật và sự thiên vị. Điều này có thể dẫn đến tổn hại cho hệ thống hoặc tổ chức.

Theo đó cần phải nhận biết các khía cạnh phát sinh tính dễ tổn thương trong hệ thống AI và có cách tiếp cận để lường trước các nguồn rủi ro, cạm bẫy, nguồn gốc các hư hỏng, nguyên nhân có tính căn nguyên, các thách thức cần phải đối mặt. Một số khía cạnh cần được đề cập là:

* Nhận biết các mối đe dọa cụ thể trong bảo mật AI: chẳng hạn như đầu độc dữ liệu, các loại hình tấn công điển hình vào dữ liệu hoặc mô hình, các phương thức đánh cắp mô hình v.v.. Ngoài ra cần phải nhận biết các mối đe dọa nhằm vào phần cứng của hệ thống.
* Các mối đe dọa đến quyền riêng tư như các hành vi thu thập, xử lý dữ liệu, truy vấn mô hình không hợp pháp, sự thiên vị trong các thuật toán ra quyết định v.v..
* Các mối đe dọa còn cần được nhận diện trong các hành vi không thể đoán trước, sự không rõ ràng (về phương pháp lập luận, suy diễn) do có những lỗ hổng trong việc thiết kế các mô hình, thuật toán sử dụng trong AI.
* Các thách thức phải đối mặt có thể làm xuất hiện các nguy cơ xuất hiện tính dễ tổn thương và các mối đe dọa trong các thành phần của hệ thống AI, chẳng hạn như hoạt động khởi đầu xây dựng các mô hình khái niệm, xác định các tham số thiết kế. Những những thách thức cũng có thể xuất phát từ bản thân các điểm yếu, nhược điểm của công nghệ, kỹ thuật áp dụng, hoạt động mô hình hóa, các quy trình huấn luyện; thiếu sót trong điều chỉnh tối ưu mô hình, dữ liệu; công tác thẩm định, đánh giá; các sai sót trong quá trình triển khai, vận hành, cách thức thu thập vào chuẩn bị dữ liệu v.v.. Do vậy tiêu chuẩn cần phải đưa ra được các chỉ dẫn để có thể thực hiện tốt các khâu, công đoạn này trong vòng đời của hệ thống AI.

Ngoài ra, tiêu chuẩn cũng cần chỉ ra các mối đe dọa phát sinh do sự hiểu biết không đầy đủ khả năng công nghệ của các hệ thống AI và việc sử dụng chúng không hợp lý bởi các bên liên quan khác nhau.

### Các biện pháp giảm thiểu

Nhận biết các tính dễ tổn thương và các mối đe dọa sẽ là không hiệu quả nếu không có các chỉ dẫn về biện pháp để giải quyết. Do vậy tiêu chuẩn cần đưa ra cách tiếp cận để giải quyết vấn đề này. Các biện pháp giảm thiểu cần được đề cập ở các vấn đề dưới đây:

* Tăng cường tính minh bạch trong hệ thống AI: được thể hiện trong sự minh bạch về thuộc tính, chức năng, mô hình, thuật toán, phương pháp huấn luyện và các bên liên quan được cung cấp các thông tin về các vấn đề trên để xem xét, đánh giá, xác định được các giá trị họ muốn thấy trong các thông tin minh bạch đó.
* Tăng cường về khả năng giải thích: chẳng hạn như diễn giải các quy trình thu thập dữ liệu, tự kiểm soát đánh giá để các bên liên quan hiểu được cách thức tạo ra kết quả, tri thức có được và sở cứ để coi kết quả là hợp lệ. Các chỉ dẫn cũng cần cung cấp cho các đối tượng áp dung các loại hình, phương pháp, mức độ giải thích cụ thể.
* Khả năng điều khiển: Cung cấp các cơ chế đáng tin cậy để người vận hành tiếp quản quyền kiểm soát đối với các hệ thống AI. Chẳng hạn như các điểm mà con người tiếp nhận điều khiển trong vòng lặp hoạt động của một quy trình nào đó thực hiện trong hệ thống.

Ngoài ra, những vấn đề dưới đây cần được xem xét một cách cụ thể về các biện pháp giảm thiêu những ảnh hưởng tiêu cực, đảm bảo hệ thống AI là đáng tin cậy phù hợp với các thuộc tính của nó:

* Chiến lược giảm sự thiên vị
* Tuân thủ quyền riêng tư
* Tăng cường sự tin cậy và khả năng phục hồi
* Giảm thiểu lỗi phần cứng
* Tăng cường sự an toàn trong vận hành, hoạt động
* Thực thi các phương pháp thẩm định, kiểm tra, đánh giá

## Lựa chọn tài liệu, tiêu chuẩn tham chiếu để xây dựng các nội dung chính của dự thảo tiêu chuẩn về tính đáng tin cậy trong AI

Hoạt động xây dựng tiêu chuẩn hóa trong lĩnh vực trí tuệ nhân tạo tại Việt Nam hiện đang ở trong giai đoạn khởi đầu do vậy hiện này chưa có kết quả nào về chuẩn hóa tính đáng tin cậy trong AI.

Để có thể xây dựng được tiêu chuẩn về tính đáng tin cậy trong AI thì cần phải lựa chọn được tài liệu, tiêu chuẩn tham chiếu để làm sở cứ xây dựng các nội dung của tiêu chuẩn này. Một số tiêu chí dưới đây cần được xem xét trong việc lựa chọn tài liệu, tiêu chuẩn tham chiếu:

* Tiêu chuẩn, tài liệu tham chiếu cần đề cập một cách toàn diện đến các khía cạnh liên quan đến về tính đáng tin cậy trong AI như đã giới thiệu ở nội dung trước. Cụ thể là tiêu chuẩn phải đề cập đến các vấn đề nhận diện, phân loại, cách tiếp cận thực thi, biện pháp giảm thiểu các khía cạnh liên quan đến tính đáng tin cậy trong các hệ thống như đã đề cập ở trên.
* Tiêu chuẩn, tài liệu làm tài liệu tham chiếu cần có các nội dung phù hợp để giải quyết các yêu cầu đề xuất cho việc xây dựng tiêu chuẩn. Các yêu cầu, chỉ dẫn phải dựa trên phương pháp tiếp cận tổng quát và toàn diện nhất; không căn cứ vào cách tiếp cận cụ thể thực thi tính đáng tin cậy trong của một loại hình sản phẩm, ứng dụng AI nhất định.
* Tiêu chuẩn, tài liệu tham chiếu đã được ban hành chính thức bởi các tổ chức tiêu chuẩn hóa quốc tế có uy tín; có tài liệu, tiêu chuẩn đã được chọn là tài liệu tham chiếu xây dựng ban hành thành TCVN trong lĩnh vực AI.

Xét theo hiệu trạng xây dựng tiêu chuẩn về tính tin cậy trong AI như đã trình bày ở nội dung trước, căn cứ vào các tiêu chí đề ra ở trên về lựa chọn tài liệu tham chiếu để xây dựng; nhóm xây dựng tiêu chuẩn nhận thấy rằng:

Tổ chức tiêu chuẩn IEEE trong thời gian qua cũng đã thực hiện một số dự án xây dựng tiêu chuẩn đề cập đến các khía cạnh thực thi tính tin cậy trong AI. Điển hình là dự án IEEE P7001 đã đưa ra các tài liệu về thực thi tính minh bạch trong các hệ thống tự trị; tiêu chuẩn IEEE 2801 – 2022 đưa ra các khuyến nghị thực hành về quản lý chất lượng các bộ dữ liệu cho AI trong lĩnh vực y tế; Tài liệu phục vụ dự án tiêu chuẩn IEEE P7000 – 2021 đưa ra tập hợp quá trình cho thiết kế hệ thống AI có tính đến các mối quan tâm về đạo đức. Tài liệu IEEE – P2937 /D6 đưa ra các chỉ dẫn về phê duyệt các điểm chuẩn đo hiệu suất hệ thống máy chủ AI. Tuy nhiên tổ chức này chưa ban hành một tài liệu, tiêu chuẩn nào đề cập một cách toàn diện và đầy đủ về các khía cạnh thực thi tính tin cậy trong các hệ thống AI. Phần lớn các tài liệu là kết quả của các dự án nói trên chưa được ban hành thành tiêu chuẩn chính thức.

Các tài liệu liên quan đến hoạt động chuẩn hóa tính tin cậy trong AI của Liên minh viễn thông quốc tế ITU-T chủ yếu đề cập đến các nhận thức, mối quan tâm trong một số khía cạnh về thực thi tính tin cậy của AI. Chẳng hạn như tài liệu DEL:AI4H do nhóm nghiên cứu trọng điểm của ITU-T về chăm sóc sức khỏe soạn thảo đề cập đến thực thi đạo đức trong AI; Tài liệu FGAI4AD-02 đề cập đến giao thức dữ liệu an toàn cho lái xe tự động. Do vậy các tài liệu được đề cập trên của tổ chức tổ chức này cũng chỉ là các thông báo kỹ thuật, chưa được ban hành thành các khuyến nghị. Và như vậy việc lựa chọn khuyến nghị hoặc tài liệu tham chiếu của ITU-T để xây dựng tiêu chuẩn về tính tin cậy trong AI là không đầy đủ.

Các tài liệu của Viện tiêu chuẩn Viễn thông châu Âu (ETSI) chủ yếu tập trung vào hoạt động được coi là thế mạnh của tổ chức này là bảo mật AI – chỉ là một trong các khía cạnh thực thi về tính tin cậy của AI. Các tài liệu điển hình của tổ chức này là: ETSI GR SAI 006 V1.1.1 (2022-03) đề cập đến vai trò của phần cứng trong bảo mật trí tuệ nhân tạo; ETSI GR SAI 002 V1.1.1 (2021-08) đề cập đến bảo mật chuỗi cung ứng dữ liệu trong AI. Như vậy có thể thấy rằng việc xem xét áp dụng các tài liệu của tổ chức này cũng là không phù hợp với yêu cầu, nội dung cho tiêu chuẩn về tính đáng tin cậy trong AI.

Hoạt động tiêu chuẩn hóa phối hợp của hai tổ chức ISO và IEC (ISO/IEC) trong lĩnh vực AI được thực hiện bởi nhóm nghiên cứu SC42 đạt được nhiều kết quả trong việc ban hành các tiêu chuẩn liên quan đến lĩnh vực AI. SC42 là nhóm nghiên cứu đã tập hợp được rất nhiều các chuyên gia từ các tổ chức tiêu chuẩn quốc tế khác và từ các lĩnh vực công nghiệp, các nhà nghiên cứu phát triển, cung cấp các sản phẩm về AI. Do vậy có thể nói đây là nhóm nghiên cứu hiệu quả nhất trong hoạt động chuẩn hóa về AI nói chung và trong lĩnh vực chuẩn hóa về tính đáng tin cậy trong AI nói riêng.

Trên cơ sở đề ra các tiêu chí về lựa chọn tài liệu tham chiếu, cũng như xem xét tình trạng ban hành, nội dung của các tiêu chuẩn, tài liệu của các tổ chức tiêu chuẩn hóa ở trên; nhóm xây dựng tiêu chuẩn quan tâm đặc biệt tới tiêu chuẩn ISO/IEC TR 24028:2020: Công nghệ thông tin – Trí tuệ nhân tạo – Tổng quan về tính đáng tin cậy trong trí tuệ nhân tạo (Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence). Trên cơ sở xem xét các nội dung tiêu chuẩn, đối chiếu với tiêu chí lựa chọn tài liệu, tiêu chuẩn tham chiếu ở trên; nhóm nghiên cứu xây dựng tiêu chuẩn nhận thấy chuẩn này đề cập mang tính toàn diện đến các vấn đề về thực thi và đảm bảo tính tin cậy trong AI, cụ thể như sau:

* Tiêu chuẩn đã đưa ra một khuôn khổ áp dụng cho thực thi tính đáng tin cậy trong hệ thống, sản phẩm và dịch vụ AI. Trong đó có cách thức nhận biết các phân lớp về độ tin cậy, xác định các tiêu chuẩn hiện có áp dụng để thực thi các chức năng tin cậy, quản lý rủi ro, các phương pháp tiếp cận quản lý rủi ro có sự trợ giúp của phần cứng.
* Tiêu chuẩn cũng xác định các loại hình bên liên quan trong quá trình phát triển, thiết kế, vận hành các hệ thống, sản phẩm, dịch vụ AI bằng việc đưa ra các khái niệm, loại hình, tài sản, giá trị của các bên liên quan trong hệ sinh thái phát triển AI
* Tiêu chuẩn đã đưa ra các định nghĩa, xác định các yêu cầu đối với các khía cạnh liên quan đến thực thi tính đáng tin cậy của AI, chẳng hạn như trách nhiệm, cách nhiệm giải trình, tính an toàn, các vấn đề rủi ro về bảo mật, vấn đề về sự thiên vị, quyền riêng tư, tính minh bạch, khả năng giải thích v.v..Ngoài ra, các nội dung trong tiêu chuẩn cũng đã cung cấp các thông tin mang tính hướng dẫn thực thi, tuân thủ các yêu cầu; chiến lược giảm thiểu rủi ro ở mọi phạm vi thực hiện có tác động đến tính đang tin cậy của hệ thống, sản phẩm AI.
* Cách tiếp cận thực hiện để đảm bảo tính đáng tin cậy trong AI dựa trên các nguyên tắc chung, mang tính tổng quát và nhất quán. Các phương pháp thực hiện được xây dựng không dựa vào các loại hình AI cụ thể nào đang có. Điều này tạo điều kiện áp dụng một cách phù hợp để phát triển các hệ thống, sản phẩm và dịch vụ AI khác nhau theo từng phân lớp hoặc phạm vi thực hiện của các đối tác liên quan.

Tiêu chuẩn ISO/IEC TR 24028:2020 đã soạn thảo từ năm 2019, dự thảo cuối cùng được công bố để lấy ý kiến biểu quyết. Tiêu chuẩn này trải qua quá trình sửa đổi, biểu quyết, phê duyệt theo quy trình phê duyệt của ISO/IEC và được thông qua ban hành chính thức thành tiêu chuẩn quốc tế vào tháng 5/2020.

Hiện tại Bộ Khoa học và Công nghệ cũng đã xây dựng một số tiêu chuẩn trong lĩnh vực AI dựa trên các tài liệu tham chiếu chính là các tiêu chuẩn do nhóm SC42 (là nhóm soạn thảo tiêu chuẩn ISO/IEC TR 24028:2020 xây dựng, cụ thể như sau:

TCVN 13238:2020: ISO/IEC 20546:2019 – Công nghệ thông tin – Dữ liệu lớn – Tổng quan và từ vựng (Information technology – Big data – Overview and vocabulary)

TCVN 13239-2:2020: ISO/IEC TR 20547-2:2018 – Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn – Phần 2: Các trường hợp sử dụng và yêu cầu dẫn xuất (Information technology – Big data reference architecture – Part 2: Use cases and derived requirements)

TCVN 13239-5:2020: ISO/IEC TR 20547-5:2018 – Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn – Phần 5: Lộ trình tiêu chuẩn (Information technology – Big data reference architecture – Part 5: Standards roadmap)

Các tiêu chuẩn tham chiếu nói trên có cùng số hiệu phân loại của ISO/IEC với tiêu chuẩn ISO/IEC TR 24028:2020 (35.020). Điều này tạo điều kiện thuận lợi hơn trong việc hướng tới xây dựng và ban hành tiêu chuẩn một cách nhất quán, hài hòa với hệ thống tiêu chuẩn quốc tế cụ thể cũng như công tác phân loại tiêu chuẩn sau này.

Như vậy tiêu chuẩn ISO/IEC TR 24028:2020 đáp ứng đầy đủ các tiêu chí cho việc lựa chọn tài liệu, tiêu chuẩn tham chiếu để xây dựng dự thảo về tính đáng tin cậy trong AI.

Với các lý do như trên, nhóm nghiên cứu đã chọn tài liệu tiêu chuẩn **ISO/IEC TR 24028:2020** làm tiêu chuẩn tham chiếu để xây dựng dự thảo tiêu chuẩn quốc gia về tính đáng tin cậy trong AI.

# Giải thích nội dung dự thảo tiêu chuẩn

## Xây dựng dự thảo tiêu chuẩn

Dự thảo được xây dựng theo hình thức chấp thuận nguyên vẹn các nội dung theo tài liệu tham chiếu được chọn và phù hợp với các tiêu chí để thực hiện mục tiêu xây dựng TCVN. Cụ thể như sau:

* Biên dịch các nội dung của tiêu chuẩn tham chiếu
* Thực hiện so sánh đối chiếu các định nghĩa, thuật ngữ bằng tiếng Anh trong tài liệu tham chiếu và các định nghĩa thuật ngữ đã dùng tại các tiêu chuẩn phù hợp được ban hành trong nước để có sự đồng nhất về các định nghĩa, thuật ngữ dịch ra tiếng Việt
* Trình bày thể thức, định dạng dự thảo tiêu chuẩn quy định trong TCVN 1-2: 2008.

## Các nội dung chính của dự thảo tiêu chuẩn

***1 Phạm vi áp dụng***

Phạm vi của tiêu chuẩn xem xét các vấn đề dưới đây liên quan đến thực thi tính đáng tin cậy trong các hệ thống AI cụ thể là:

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

Tiêu chuẩn không đưa ra các quy định về mức độ mang tính định lượng để đánh giá các đặc tả về tính đáng tin cậy của các hệ thống AI

Các bên liên qua áp dụng tiêu chuẩn cho các hoạt động nghiên cứu, phát triển, thiết kế, triển khai, vận hành, sử dụng và quản lý các sản phẩm, dịch vụ và hệ thống AI. Do vậy đối tượng sử dụng tiêu chuẩn là các cơ quan quản lý nhà nước, doanh nghiệp, tổ chức và cá nhân quan tâm đến các vấn đề thực thi tính tin cậy của AI trong các hoạt động nói trên

***2. Tài liệu viện dẫn***

Các tài liệu tham chiếu cho các nội dung của dự thảo tiêu chuẩn

***3. Các thuật ngữ và định nghĩa***

Nội dung này đưa ra thuật ngữ và định nghĩa được sử dụng trong tiêu chuẩn này.

***4 Tổng quan***

Nội dung này đưa ra cái nhìn tổng quát nhất về tiêu chuẩn; giúp cho người sử dụng tiêu chuẩn hiểu cách tiếp cận xây dựng bố cục, nội dung có trong tiêu chuẩn

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

Nội dung đề cập đến việc áp dụng các phương pháp đã có (được quy định trong các tiêu chuẩn được ban hành) để nhận biết và giải quyết các mối quan tâm về tính đáng tin cậy của các hệ thống AI ở các khía cạnh khác nhau (chẳng hạn như đối với tính đáng tin cậy của hệ thống phần mềm, phần cứng, các quy trình quản lý chất lượng sản phẩm). Trên cơ sở đó tiêu chuẩn khuyến nghị các phương pháp mới, hoặc các cải tiến cho phương pháp hiện có phù hợp áp dụng cho tạo ra một khuôn khổ để thực thi tính tin cậy của AI.

Theo đó các nội dung phần này đề cập đến các vấn đề dưới đây

5.1 Cơ sở

5.2 Nhận biết các lớp tin cậy

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

5.4 Áp dụng quản lý rủi ro

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

Các chỉ dẫn có trong nọi dung này giúp các đối tác liên quan xây dựng khung công việc, tiêu chuẩn áp dụng về thực thi tính đáng tin cậy trong hệ thống AI trong các hoạt động thiết kế, vận hành, quản lý các hệ thống, sản phẩm, dịch vụ AI.

**6. Các bên liên quan**

Nội dung này thực hiện phân loại các loại hình bên liên quan, xác định các loại hình tài sản, giá trị, cũng như mối quan hệ, ảnh hưởng, tác động qua lại giữa các đối tượng này trong hệ sinh thái AI. Cụ thể nó là rõ một số vấn đề như:

* Ảnh hưởng của yếu tố công nghệ đối với các bên liên quan xét trong một hệ thống, sản phẩm, dịch vụ AI;
* Xác định các loại hình tài sản và giá trị trong môi trường phát triển, ứng dụng và sử dụng AI

Trên cơ sở đó các vấn đề dưới đây được đề cập trong nội dung này:

6.1 Các khái niệm chung

6.2 Loại hình

6.3 Tài sản

6.4 Giá trị

Nội dung giúp các các đối tượng áp dụng tiêu chuẩn nhận thức rõ vị trí, vai trò, nhận biết các giá trị, tài sản mà họ nắm giữ để từ đó có những hành động phù hợp trong hoạt động của mình (chẳng hạn như nghiên cứu phát triển, vận hành, sử dụng, quản lý v.v..).

***7 Nhận biết các mối quan tâm cấp cao***

Các chỉ dẫn có trong nội dung này đề cập đến vai trò của tổ chức (các bên liên quan) trong việc thực thi hoặc quản lý các vấn đề về tính tin cậy của các hệ thống AI. Có hai vấn đề chính được đề cập đó là trách nhiệm giải trình giữa các bên liên quan và đảm bảo tính an toàn của các hệ thống AI. Các tác hại gây ra bởi sự không an toàn không chi là tác hại hữu hình (ví dụ, đối với sức khỏe của sinh vật, tài sản và môi trường vật chất) mà còn cả tác hại vô hình (ví dụ đối với môi trường xã hội và văn hóa). Các nội dung của phần này bao gồm:

7.1 Trách nhiệm, trách nhiệm giải trình và quản trị

7.2 Tính an toàn

Nội dung này giúp cho các đối tượng, bao gồm cả nhà cung cấp, người sử dụng và nhà quản lý xem xét áp dụng tiêu chuẩn để thực thi tính đáng tin cậy của AI trong phạm vi một tổ chức, doanh nghiệp, cơ quan quản lý.

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

Nội dung đưa ra các chỉ dẫn để nhận biết các khía cạnh phát sinh tính dễ tổn thương trong hệ thống AI cũng như cách tiếp cận để lường trước các nguồn rủi ro, cạm bẫy, nguồn gốc các hư hỏng, các thách thức cần phải đối mặt trong các hoạt động thực thi. Để cụ thể hóa, nội dung đề cập đến các vấn đề cụ thể dưới đây:

8.1 Yêu cầu chung

8.2 Các mối đe dọa bảo mật cụ thể của AI (đầu độc dữ liệu, tấn công đối nghịch, đánh cắp mô hình, mối đe dọa nhằm vào phần cứng).

8.3 Các mối đe dọa quyền riêng tư điển hình trong AI (thu thập xử lý dữ liệu truy vấn mô hình

8.4 Thiên vị

8.5 Tính không thể đoán trước

8.6 Tính không rõ ràng

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

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

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

khó đoán, không rõ ràng, điểm yếu kỹ thuật v.v..).

Nội dung này giúp các đối tượng sử dụng tiêu chuẩn hiểu rõ tính dễ bị tổn thương, các mối đe dọa và thách thức gặp phải trong quá trình triển khai, vận hành, sử dụng và quản lý các hệ thống AI để từ đó có những hành động phù hợp cho các hoạt động nói trên.

***9. Các biện pháp giảm thiểu***

Trên cơ sở nhận biết, xác định tính dễ bị tổn thương, các mối đe dọa, các thách thức gặp phải đối với các hệ thống AI như đề cập ở trên; nội dung này đưa ra cách tiếp cận để giải quyết vấn đề đó. Chỉ dẫn về thực thi các biện pháp giảm thiểu được thể hiện ở các nội dung dưới đây:

9.1 Yêu cầu chung

9.2 Tính minh bạch

9.3 Khả năng giải thích

9.4 Khả năng điều khiển

9.5 Các chiến lược giảm thiểu tính thiên vị

9.6 Quyền riêng tư

9.7 Tính tin cậy, khả năng phục hồi và tính mạnh mẽ

9.8 Giảm thiểu lỗi phần cứng

9.9 Tính an toàn trong hoạt động

9.10 Kiểm tra và đánh giá

9.11 Sử dụng và khả năng sử dụng

Các nội dung trên giúp cho các đối tượng áp dụng tiêu chuẩn nắm được các biện pháp giảm thiểu nhằm giải quyết, khắc phục tính dễ tổn thương; biện pháp đề phòng các mối đe dọa và giải quyết các thách thức gặp phải trong hoạt động của các hệ thống AI.

Đặc biệt các nội dung chỉ dẫn về kiểm tra và đánh giá; sử dụng và khả năng sử dụng là rất hữu ích cho các cơ quan quản lý trong việc đề xuất các quy định về giới hạn phạm vi sử dụng, xác minh, thẩm định phục vụ việc công tác kiểm tra, đánh giá, công nhận hợp chuẩn đối với các hệ thống, dịch vụ, sản phẩm AI sau này.

***10. Kết luận***

Tổng kết lại các nội dung được trình bày trong tiêu chuẩn. Nội dung kết luận nhấn mạnh ý nghĩa, giá trị tầm quan trọng của việc thực thi tính đáng tin cậy trong AI thông qua các yêu cầu, chỉ dẫn được trình bày trong tiêu chuẩn.

***Phụ lục (tham khảo)***

Cung cấp các thông tin, kết quả nghiên cứu sự liên quan, ảnh hưởng của các yếu tố về thực thi tính đáng tin cậy trong AI đối với các vấn đề xã hội

Bảng 1 – Bảng đối chiếu nội dung dự thảo tiêu chuẩn với tài liệu tiêu chuẩn tham chiếu ISO/IEC 24028:2020

|  |  |  |
| --- | --- | --- |
| **Dự thảo tiêu chuẩn xxxxx:2022: Công nghệ thông tin – Trí tuệ nhân tạo – Tổng quan về tính đáng tin cậy trong Trí tuệ nhân tạo** | **Tài liệu tham chiếu**  **(ISO/IEC 24028:2020)** | **Sửa đổi, bổ sung** |
| **1 Phạm vi áp dụng** | **1 Scope** | **Chấp thuận nguyên vẹn** |
| **2 Tài liệu viện dẫn** | **2 Normative references** | **Chấp thuận nguyên vẹn** |
| **3 Các thuật ngữ và định nghĩa** | **3 Terms and definitions** | **Chấp thuận nguyên vẹn** |
| **4 Tổng quan** | **4 Overview** | **Chấp thuận nguyên vẹn** |
| **5 Các khuôn khổ hiện có áp dụng cho tính đáng tin cậy** | **5 Existing frameworks applicable to trustworthiness** | **Chấp thuận nguyên vẹn** |
| 5.1 Cơ sở | 5.1 Background | Chấp thuận nguyên vẹn |
| 5.2 Nhận biết các lớp tin cậy | 5.2 Recognition of layers of trust | Chấp thuận nguyên vẹn |
| 5.3 Áp dụng các tiêu chuẩn chất lượng phần mềm và dữ liệu | 5.3 Application of software and data quality standards | Chấp thuận nguyên vẹn |
| 5.4 Áp dụng quản lý rủi ro | 5.4 Application of risk management | Chấp thuận nguyên vẹn |
| 5.5 Các phương pháp tiếp cận có sự trợ giúp của phần cứng | 5.5 Hardware-assisted approaches | Chấp thuận nguyên vẹn |
| **6 Các bên liên quan** | **6 Stakeholders** |  |
| 6.1 Khái niệm chung | 6.1 General concepts | Chấp thuận nguyên vẹn |
| 6.2 Loại hình | 6.2 Types | Chấp thuận nguyên vẹn |
| 6.3 Tài sản | 6.3 Assets | Chấp thuận nguyên vẹn |
| 6.4 Giá trị | 6.4 Values | Chấp thuận nguyên vẹn |
| **7 Nhận biết các mối quan tâm cấp cao** | **7 Recognition of high-level concerns** |  |
| 7.1 Trách nhiệm, trách nhiệm giải trình và quản trị | 7.1 Responsibility, accountability and governance | Chấp thuận nguyên vẹn |
| 7.3 Tính an toàn | 7.2 Safety | Chấp thuận nguyên vẹn |
| **8 Tính dễ bị tổn thương, các mối đe dọa và thách thức** | 8 Vulnerabilities, threats and challenges |  |
| 8.1 Yêu cầu chung | 8.1 General | Chấp thuận nguyên vẹn |
| 8.2 Các mối đe dọa bảo mật cụ thể của AI | 8.2 AI specific security threats | Chấp thuận nguyên vẹn |
| 8.3 Các môi đe dọa quyền riêng tư điển hình trong AI | 8.3 AI specific privacy threats | Chấp thuận nguyên vẹn |
| 8.4 Thiên vị | 8.4 Bias | Chấp thuận nguyên vẹn |
| 8.5 Tính không thể đoán trước | 8.5 Unpredictability | Chấp thuận nguyên vẹn |
| 8.6 Tính không rõ ràng | 8.6 Opaqueness | Chấp thuận nguyên vẹn |
| 8.7 Những thách thức liên quan đến đặc điểm kỹ thuật của hệ thống AI | 8.7 Challenges related to the specification of AI systems | Chấp thuận nguyên vẹn |
| 8.8 Những thách thức liên quan đến triển khai các hệ thống AI | 8.8 Challenges related to the implementation of AI systems | Chấp thuận nguyên vẹn |
| 8.9 Những thách thức liên quan đến sử dụng các hệ thống AI | 8.9 Challenges related to the use of AI systems | Chấp thuận nguyên vẹn |
| 8.10 Lỗi phần cứng hệ thống | 8.10 System hardware faults | Chấp thuận nguyên vẹn |
| **9 Các biện pháp giảm thiểu** | **9 Mitigation measures** |  |
| 9.1 Yêu cầu chung | 9.1 General | Chấp thuận nguyên vẹn |
| 9.2 Tính minh bạch | 9.2 Transparency | Chấp thuận nguyên vẹn |
| 9.3 Khả năng giải thích | 9.3 Explainability | Chấp thuận nguyên vẹn |
| 9.4 Khả năng điều khiển | 9.4 Controllability | Chấp thuận nguyên vẹn |
| 9.5 Các chiến lược giảm tính thiên vị | 9.5 Strategies for reducing bias | Chấp thuận nguyên vẹn |
| 9.6 Quyền riêng tư | 9.6 Privacy | Chấp thuận nguyên vẹn |
| 9.7 Tính tin cậy, khả năng phục hồi và tính mạnh mẽ | 9.7 Reliability, resilience and robustness | Chấp thuận nguyên vẹn |
| 9.8 Giảm thiểu lỗi phần cứng hệ thống | 9.8 Mitigating system hardware faults | Chấp thuận nguyên vẹn |
| 9.9 Tính an toàn trong hoạt động | 9.9 Functional safety | Chấp thuận nguyên vẹn |
| 9.10 Kiểm tra và đánh giá | 9.10 Testing and evaluation | Chấp thuận nguyên vẹn |
| 9.11 Sử dụng và khả năng sử dụng | 9.11 Use and applicability | Chấp thuận nguyên vẹn |
| **10 Kết luận** | **10 Conclusions** | **Chấp thuận nguyên vẹn** |
| **Phụ lục A (tham khảo)** | **Annex A ((informative)** | **Chấp thuận nguyên vẹn** |

# Khuyến nghị áp dụng tiêu chuẩn

Một số khuyến nghị dưới đây về áp dụng tiêu chuẩn trong thực tế:

***Mục tiêu chuẩn hóa:***

Tiêu chuẩn đề cập đến các vấn đề thực thi tính đáng tin cậy trong các hệ thống AI. Nội dung có trong tiêu chuẩn làm rõ các đặc điểm, thuộc tính, xác định các đối tượng chịu tác động và ảnh hưởng tới tính đáng tin cậy trong các hệ thống AI. Các yêu cầu, chỉ dẫn đưa ra trong tiêu chuẩn đề cập đến cách thức kiểm tra, phát hiện để từ đó đưa ra các biện phạm giảm thiểu nguy cơ, rủi ro, các thách thức đối với việc tăng cường tính đáng tin cậy trong các hệ thống AI. Do vậy dự thảo tiêu chuẩn đáp ứng mục tiêu xác định chức năng, công dụng, chất lượng: xác định các công nghệ, đặc điểm kỹ thuật ở mức tiên tiến và phù hợp để áp dụng tại Việt Nam. Ngoài ra các nội dung trong tiêu chuẩn, chẳng hạn như phân loại bên liên quan, tài sản, giá trị, hoạt động xác minh, thẩm định tính đang tinh cậy trong AI cũng giúp cho các cơ quan quản lý nhận dạng các đối tượng, các đặc điểm kỹ thuật cần thiết có ảnh hưởng tới các bên liên quan trong xã hội để có những quy định, hoạt động triển khai quản lý phù hợp.

***Đối với các đối tượng áp dụng tiêu chuẩn:***

Các đối tượng áp dụng tiêu chuẩn này là các cá nhân, tổ chức, doanh nghiệp, cơ quan quản lý nhà nước hoạt động trong mọi lĩnh vực liên quan tới AI như nghiên cứu kỹ thuật công nghệ, thiết kế, phát triển, sản xuất, triển khai và quản lý nhà nước đối với các hệ thống, sản phẩm, dịch vụ dựa trên công nghệ AI (hoặc một phần dựa trên công nghệ AI). Mỗi một đối tượng áp dụng tiêu chuẩn có thể tham gia vào một hoặc nhiều khâu trong toàn bộ quá trình vòng đời của một hệ thống, sản phẩm AI. Do đó các đối tượng áp dụng cần phải có những nhận thức, hiểu biết rõ ràng và nhất quán về khuôn khổ thực thi, vị trí, vai trò, chức năng; nhận biết các thuộc tính, yêu cầu về tính đáng tin cậy; các mối quan tâm, đe dọa, thách thức trong việc thực thi và duy trì tính tin cậy của các hệ thống AI trong suốt vòng đời của nó. Các đối tượng chịu tác dộng của tiêu chuẩn ở những khía cạnh dưới đây:

* *Đối tượng tham gia thiết kế, phát triển triển khai các sản phẩm, dịch vụ AI:* Tiêu chuẩn giúp cho các bên tham gia hiểu rõ bản chất, các yêu cầu về thực thi tính đáng tin cậy để từ đó áp dụng trong việc xây dựng các mô hình khái niệm, thiết kế chức năng, biện pháp và kỹ thuật cụ thể áp dụng để đảm bảo tính đáng tin cậy của hệ thống AI theo yêu cầu. Thông qua các khuôn khổ chung về đảm bảo tính đang tin cậy, các bên liên quan xác định vị trí vai trò, để từ đó có một giao diện thống nhất trao đổi thông tin giữa các bên về phạm vi trách nhiệm thực thi trong một quy trình tổng thể đảm bảo tính đáng tin cậy của một hệ thống AI.
* *Các đối tượng vận hành, sử dụng hệ thống, sản phẩm, dịch vụ AI:* Tiêu chuẩn giúp các đối tượng này hiểu rõ các thuộc tính, cách thức hoạt động của hệ thống AI để thể hiện tính đáng tin cậy của nó. Các vấn đề như tài sản, giá trị, bản chất cơ chế hoạt động, chức năng giải thích của hệ thống AI giúp các đối tượng vận hành, sử dụng nhận thức những mối đe dọa, rủi ro tiềm tàng để từ đó có cách thức vận hành, sử dụng các hệ thống, sản phẩm, dịch vụ AI một cách đúng đắn và phù hợp.
* *Các đối tượng tham gia quản lý:* Các nội dung chỉ dẫn có trong tiêu chuẩn về kiểm tra và đánh giá; sử dụng và khả năng sử dụng là rất hữu ích cho các cơ quan quản lý trong việc đề xuất các quy định về giới hạn phạm vi sử dụng, hoạt động xác minh, thẩm định phục vụ việc công tác kiểm tra, đánh giá, công nhận hợp chuẩn đối với các hệ thống, dịch vụ, sản phẩm AI.

***Đối với hoạt động xây dựng các tiêu chuẩn trong lĩnh vực AI:***

* Tiêu chuẩn đưa ra khuôn khổ, yêu cầu và chỉ dẫn chung về thực thi tính đáng tin cậy trong các hệ thống AI. Do đó tiêu chuẩn này là cơ sở để xây dựng các tiêu chuẩn sau này về tính đáng tin cậy của hệ thống AI ứng dụng trong các lĩnh vực cụ thể, chẳng hạn như y tế, người máy, lái xe tự động v.v...
* Tiêu chuẩn áp dụng tài liệu tham chiếu là tiêu chuẩn quốc tế của tổ chức ISO/IEC theo hình thức chấp thuận nguyên vẹn. Do vậy tiêu chuẩn đáp ứng sự hài hòa đối với hệ thống tiêu chuẩn quốc tế, góp phần thúc đẩy chủ trương phát triển hội nhập trong giai đoạn hiện nay.

# Kiến nghị về hướng phát triển tiếp theo

Dự thảo tiêu chuẩn TCVN xxxxx:2022 “Công nghệ thông tin – Trí tuệ nhân tạo – Tổng quan về tính đáng tin cậy trong Trí tuệ nhân tạo” được ban hành thành tiêu chuẩn quốc gia sẽ là một trong những tiêu chuẩn đầu tiên trong lĩnh vực AI trong nước. Dự kiến trong thời gian tới cần tiếp tục triển khai hoạt động chuẩn hóa AI trong lĩnh vực thực thi tính đáng tin cậy của AI, trọng tâm hướng tới các vấn đề dưới đây:

* Tính thiên vị và ra quyết định được trợ giúp bởi AI
* Tổng quan về đạo đức và mối quan tâm về xã hội đối với AI
* Cập nhật tình hình tiêu chuẩn hóa của các tổ chức tiêu chuẩn hóa lớn trên thế giới về tính đáng tin cậy trong các lĩnh vực AI cho y tế, xe tự lái, người máy. Các hệ thống tự trị v.v..

------------------------------------------------------