AI/ML 利用システムのセキュリティ論証に必要な品質特性に関する考察

溝口 誠一郎[1]　櫻井 幸一[2]

概要：自動運転車は AI/ML 技術を利用しているが，これに伴いセーフティやセキュリティのリスクが生じる．AI の品質，特にデータの品質は，これらのリスク軽減に不可欠である．自動車分野の AI 機能安全には ISO/PAS 8800 があり，一般的な AI 利用におけるデータ品質管理には ISO/IEC 5259 がある．AI セキュリティの観点では，データの信頼性や完全性の確保が，AI の誤動作防止や悪意ある攻撃からの保護に直結するため，これらの品質特性が極めて重要である．
本稿では，これらの AI/ML の品質特性と AI セキュリティの関連について考察する．

キーワード：AI/ML, Safety, Security, Assurance

Considerations on Quality Attributes Necessary for Security Assurance of AI/ML-driven Systems

Seiichiro Mizoguchi[1]　Kouichi Sakurai[2]

Abstract: Autonomous vehicles leverage AI/ML technologies, which introduce inherent safety and security risks. AI quality, particularly data quality, is crucial for mitigating these risks. In the automotive sector, ISO/PAS 8800 addresses AI functional safety, while ISO/IEC 5259 focuses on general data quality management for AI applications. From an AI security perspective, ensuring data reliability and integrity directly contributes to preventing AI malfunctions and protecting against malicious attacks, making these quality attributes critically important.
This paper will explore the connection between these AI/ML quality attributes and AI security.

Keywords: AI/ML, Safety, Security, Assurance

1. はじめに

自動運転車は，AI/ML 技術を駆使して動作する．車両に搭載されたカメラや LiDAR，レーダー等の各種センサーから得られる膨大なデータを，AI がリアルタイムに解析し，歩行者や他の車両，道路標識などを識別し，その動向を予測することで，周囲の環境を認識する役割を担っている．AI/ML は，多岐にわたる交通状況，天候，道路種別等の走行データを学習することにより，状況に応じた適切な判断を下す能力を獲得する．具体的には，信号の赤点灯時の停止や車線変更といった，複雑な運転操作に関する判断の精度を，学習を通じて向上させる．テスラの「オートパイロット」や Google の「Waymo」といったシステムは，この AI/ML によって収集・学習されたデータに基づき，運転操作の大部分を自動で実行する．これらの技術の応用は，人為的な運転ミスを低減し，交通安全性の向上に寄与すると考えられている．

一方，自動運転車の安全性については，いくつかの事例で問題が指摘されている．2018 年には，アリゾナ州で Uber の自動運転試験車が歩行者をはねて死亡させる事故が発生した[1]．また，テスラの「オートパイロット」使用中に，障害物認識の不備が原因で衝突事故に至るケースも報告されている[2]．これらの事例は，AI/ML の判断能力やセンサーの限界，システムの過信といった課題を浮き彫りにし，技術の発展と安全性の確保とのバランスが重要であることを示している．

自動運転車は AI/ML により周囲を認識するが，AI 特有のセキュリティ脆弱性が存在する．例えば，**敵対的サンプル**と呼ばれる，人間には無害なわずかな改変をデータに加える手法がある[3]．道路標識に特殊なステッカーを貼ることで，AI がこれを誤認識し，重大な事故を引き起こすリスクが指摘されている．また，顔認証システムも，特殊な模様のメガネ等により欺かれる検証事例がある．このように，従来のサイバーセキュリティ対策に加え，AI の誤認識を誘発する攻撃にも対処する必要があり，AI が社会インフラに普及する上で，その安全性確保は喫緊の課題である．

そういった中，AI システムの安全性と信頼性を確保するため，複数の国際標準やガイドラインが発行されている．

1 DNV ビジネスアシュアランスジャパン株式会社
DNV Business Assurance Japan K.K
2 九州大学
Kyushu University

これらは，AI 特有のリスクを管理することを目的としている．

ISO/IEC 42001 - Information technology — Artificial intelligence — Management system [4]は，組織が AI システムを適切に管理するためのマネジメントシステムに関する規格である．これは，情報セキュリティの ISO 27001 - Information security, cybersecurity and privacy protection — Information security management systems — Requirements [5]と同様に，リスクベースで AI の倫理，説明責任，透明性，そしてセキュリティを体系的に管理する枠組みを提供する．また，ISO/PAS 8800 - Road vehicles — Safety and artificial intelligence [6]は，特に自動運転車のような機能安全が求められる分野における AI の安全性を扱う．AI が予期せぬ振る舞いや誤認識を起こすリスクを評価し，その対策を講じるためのガイダンスを定めている．さらに，AI の性能に不可欠なデータの品質については，ISO/IEC 5259 - Artificial intelligence — Data quality for analytics and machine learning (ML) [7]シリーズが基準を設けている．データ収集から利用に至るまでのライフサイクル全体にわたる品質管理を規定することで，データに起因するバイアスや不具合を防ぎ，AI システムの信頼性を高めることを目指す．これらの標準は，AI を社会インフラに組み込む上でのセーフティネットの役割を担っている．

UNECE WP29 では，2025 年 3 月に，人工知能に関する非公式作業部会（Informal Working Group on Artificial Intelligence）の設立活動がすすめられている[8]．それに先立ち，Proposal for a draft resolution with guidance on Artificial Intelligence in the context of road vehicles[9]では，”第 3 項: ソフトウェアに組み込まれた AI システムは，訓練された後，権限のある関係者や認証プロセスによって検証されることが推奨される .この検証では，安全性，セキュリティ，環境性能，およびその他の関連要件が評価されるべきである．”という提案がなされており，AI システムのセキュリティの評価が，認証プロセスにおいて実施される可能性を示唆している．

このように，自動車分野に限らず，AI/ML を利用するシーンでは，AI 利用におけるリスク管理に関するガイダンスが発行されており，リスクを適切にコントロールする体制が求められつつある．本稿では，特に AI システムのセキュリティに関連する AI システムの品質特性について調査結果をまとめる．

2. 国際規格について

あらためて，各国際規格の位置づけについて確認する．

2.1 ISO/IEC 38507

ISO/IEC 38507 - Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations [10]は，人工知能（AI）のガバナンスに関する国際規格である．

組織が，AI のガバナンスを効果的に行うためにいくつかの原則を示しているが，その中に，リスクマネジメントの要件がある．ここでのリスクは，組織が，AI システムを導入する際に，組織として維持しなければならない性質を脅かすものをいう．組織が維持すべき性質として以下が挙げられている：

- 説明責任と責任
- 評判と信頼
- 注意義務（内部および外部の利害関係者に対する）
- 安全性（物理的または感情的なもの）
- セキュリティとプライバシー
- データセキュリティ
- 透明性

自動運転などの，利用者の安全に関わる部分で用いられる AI システムは，当たり前ではあるが，この安全性の原則を守らなければならない．また，AI システムを開発するにあたっては，AI モデルや学習・推論データなどの，データの保護についても考慮する必要がある．

2.2 ISO/IEC 42001

ISO/IEC 42001 は，AI マネジメントシステムと呼ばれる，AI システムを利用する組織が，その AI システムの利活用におけるリスク管理を行うための実践的な枠組みを規定するものである．

組織は，アシロマ AI 原則[11]や，ISO/IEC 38507 に規定されるような原則が守られるよう，組織の活動を制御する必要がある．そのような AI 利用における原則と，組織が本来達成しなければならない目的を阻害するようなリスクを定義し，一般的なリスクマネジメントの枠組みに則って，AI システムのリスク管理を行う．

本規格の特徴として，AI 技術者目線のリスク管理方針も規定されている．AI システムのライフサイクルを定義して，開発から保守運用までの中で，品質管理することの重要性を述べている．

ISO/IEC 42001 では，セキュリティとプライバシーにも触れられており，AI マネジメントシステムが達成すべき目的を阻害する要因としてこれらが挙げられている．情報セキュリティに関しては，ISO/IEC 42001 と ISO/IEC 27001 が独立して組織内に存在する必要は無く，むしろ統合できることが主張されている．

2.3 ISO/PAS 8800

自動車業界における AI 利用の安全関連規格として，ISO/PAS 8800[6]が 2024 年 12 月 13 日に発行された．ISO/PAS 8800 は，ISO/IEC TR 5469 - Artificial intelligence — Functional safety and AI systems [12]をベースに，AI 利用システムにおける安全に関する特性やテスト戦略，AI システムの開発およびアーキテクチャの冗長性パターンなど，AI システムの

機能安全に関する一般的なフレームワークを，自動車に適用したものである．自動車分野の安全管理については，ISO 26262 - Road vehicles — Functional safety [13]や ISO21448 - Road vehicles — Safety of the intended functionality [14]といった先行規格があり，システムに AI コンポーネントを含む場合に，ISO/PAS 8800 が適用される想定となっている．

ISO/PAS 8800 中で，自動車で用いられる AI システムに割り当てられる安全要件には，敵対的攻撃やデータポイズニングに対する堅牢性といった，AI 特有の特性が含まれている．

2.4 ISO/IEC 5259 シリーズ

ISO/IEC 5259[7]は，AI/ML で処理されるデータそのものの品質管理に関する規格である．

ISO/IEC 5259 では，データ品質を，データが指定されたコンテキストに対する組織の要件を満たすというデータの特性であると定義しており，それは分析結果の品質や ML モデルの性能に影響を与えるとしている．

データ品質のガバナンスのために，データ品質原則を用意し，データライフサイクルモデル (Data Life Cycle Model, DLC Model) に沿って，データ品質の測定と，データ品質の評価を行うという構造になっている．

ISO/IEC 5259 の中で，データセキュリティに関する言及がある．それは，”データセットは，権限を持つ担当者やプロセスが利用できるよう，また不適切な改ざんを防ぐため，DLC モデルの全段階を通じて安全に保たれるべきである．データセットへの不適切な変更は，それ自体が機械学習モデルや他の分析タスクから誤った結果を引き起こす可能性がある．”という内容である．DLC では，Data Preparation の段階で，集めたデータに対して，データセキュリティとデータプライバシーが考慮されているかどうかを確認すべきとしており，参考として，ISO/IEC 27001 や，ISO/IEC 27701 - Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines [15]が参照されている．

2.5 Proposal for a draft resolution with guidance on Artificial Intelligence in the context of road vehicles の内容

本文書[9]では，AI-based システムが，既存の Cybersecurity と Software Update の要件に与える影響を示唆している．Annex 1 の冒頭には以下のような記述がある：

- 自動車における AI の特性と課題: 自動車製品で使用される AI ベースのシステムは，安全性とセキュリティを確保しつつ，モデルのドリフト（drift）や陳腐化（staleness），モデルの複雑さ，頑健性（robustness），検証可能性（verifiability），予測可能性（predictability），過学習（overfitting）といった望ましい特性のバランスを取ることが求められる ．また，AI ベースのシステムは，システムアップデートの可能性を提供すべきである ．
- ソフトウェアアップデートに関する再評価の必要性: サイバーセキュリティとソフトウェアアップデートに関する既存の規定が，AI ベースのシステムのアップデートに適切に対応しているかを確認するため，定期的な再評価が必要になる可能性がある ．

通常の Cybersecurity 上の脅威が，これら AI の特性を侵害する脅威になることを，AI リスクマネジメントの中で意識する必要がある．

3. AI/ML の品質特性とセキュリティ特性の関係についての課題と考察

これまで，UN-R155[16]や，ISO/SAE 21434 - Road vehicles — Cybersecurity engineering [17]で扱われていたリスクは，基本的には，道路利用者の SFOP (Safety, Financial, Operational and Privacy) に影響を及ぼすものを中心に管理していた．しかし，AI システムにおいては，AI 利用において，組織が維持すべき原則に対する対応の説明性が求められるようになる．ISO/SAE 21434 では，安全に影響を及ぼすシステム上の振る舞いと，システム上のアセットのセキュリティ特性(例として，機密性，完全性，可用性，あるいは真正性，等)の侵害を紐づけるにとどまっていたが，AI/ML 利用においては，組織が AI/ML 利用するにあたって AI 原則を満たしていることを，それらの原則を満たせなくなるようなリスクが組織的に管理されていることを併せて説明しなければならない．

自動車分野特有の課題としては，分散開発を基本として組織構造になっており，そこに AI ガバナンスを適用する必要があるということが挙げられる．UN-R155 などの発行により，自動車分野において，サイバーセキュリティの重要性は認識されたが，ISO/SAE 21434 に基づくプロセスの構築が優先的に行われたため，本来開発のプロセスと，ある意味独立した形になっている企業は多い．セキュリティ技術者もそもそも少ないため，製品セキュリティを統括するチームに，製品セキュリティに関する説明が一任される傾向にあるが，製品のセキュリティポリシーは，その製品の目的や，目的を達成するための使われ方に依存するため，Security by Design の考え方で，本来開発の上流からしっかりと検討することが望ましい．例えば，ISO/SAE 21434 では，上流の Concept と，下流の Development を繋ぐ最上位の要求として，Cybersecurity Goal が定義されており，資産のセキュリティ特性（CIA など）を担保することが，Goal の典型的な表現となっている．そういった中で，AI/ML 利用上の原則に対して，セキュリティコントロールがどのように行われているのかを説明できる必要がある．

4. おわりに

AI/ML 利用にあたり，組織やシステムそのものに対する要件が述べられた国際規格やガイダンスが発行されている．組織としては，これらの規格から，自身に必要な要件を選択し，それを管理していく体制が必要である．セキュリティの観点では，AI システム自身のセキュリティ確保の観点と，企業が AI システムを利用するにあたっての，運用上のセキュリティ確保の両面があるため，AI の品質特性が，セキュリティ上の特性としてどのように関連するかを継続して議論する必要があるため，引き続き調査を継続する．

参考文献

[1] How a Self-Driving Uber Killed a Pedestrian in Arizona, The New York Times, MARCH 21, 2018, https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

[2] 米テスラ，「自動運転モード」作動中に初の死亡事故，日本経済新聞, 2016-07-01, https://www.nikkei.com/article/DGXLASGN01H0T_R00C16A7000000/, (参照 2025-08-22)

[3] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, “EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES”, arXiv:1412.6572v3, 20 Mar. 2015, https://arxiv.org/pdf/1412.6572

[4] ISO/IEC 42001:2023 – AI Management systems

[5] ISO/IEC 27001:2022 - Information security, cybersecurity and privacy protection — Information security management systems — Requirements

[6] ISO/PAS 8800:2024 - Road vehicles — Safety and artificial intelligence

[7] ISO/IEC 5259 - Artificial intelligence — Data quality for analytics and machine learning (ML)

[8] Artificial Intelligence Informal Working Group – justification and scope, WP.29-195-20, 195th Wp.29, 4-7 March 2025, https://unece.org/transport/documents/2025/03/informal-documents/united-kingdom-artificial-intelligence-informal

[9] Proposal for a draft resolution with guidance on Artificial Intelligence in the context of road vehicles, 21 June 2024, https://unece.org/transport/documents/2024/06/working-documents/secretariat-proposal-draft-resolution-guidance

[10] ISO/IEC 38507 - Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

[11] アシロマの原則, 2 August 2017, https://futureoflife.org/open-letter/ai-principles-japanese/

[12] ISO/IEC TR 5469:2024 - Artificial intelligence — Functional safety and AI systems

[13] ISO 26262 - Road vehicles — Functional safety

[14] ISO 21448:2022 - Road vehicles — Safety of the intended functionality

[15] ISO/IEC 27701:2019 - Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines

[16] UN Regulation No. 155 - Cyber security and cyber security management system

[17] ISO/SAE 21434:2021 - Road vehicles — Cybersecurity engineering