AI を利用する車両システムのセキュリティと安全論証について

溝口 誠一郎[1]　櫻井 幸一[2]

概要：自動運転の普及に向けての課題として，AI を利用するシステムのサイバーセキュリティの論証と対策である．AI に対する攻撃と予防的対策については研究が進んでいるが，攻撃を受けたときの対策については課題がある．本発表では，AI および AI を利用するシステムに対するセキュリティの脅威と対策について，機能安全的側面から考察する．

キーワード：Vehicle Security, AI, Assurance Case

Cybersecurity and Safety Argument of Vehicle Systems using AI

Seiichiro Mizoguchi[1]　Kouichi Sakurai[2]

Abstract: One of the challenges for the widespread adoption of autonomous driving is the demonstration and countermeasures of cybersecurity for systems that utilize AI. While research on attacks against AI and preventive measures is advancing, there are challenges in responding to attacks when they occur. This presentation will consider the security threats and countermeasures against AI and systems utilizing AI from the perspective of functional safety.

Keywords: Vehicle Security, AI, Assurance Case

1. はじめに

製品のサイバーセキュリティが維持管理されていることを説明することが求められている．自動車においては，UN-R155[1]により，車両がサイバーセキュリティの脅威に対して対策が取られていることを説明する必要がある．UN-R155 は，日本における道路運送車両の保安基準[2]として採用されている．また，欧州 Cyber Resilience Act[3]により，欧州で流通する製品については，自動車と同様に，そのサイバーセキュリティが維持されていることを説明しなければならない．CRA は EU 圏が対象であるが，日本においては経済産業省が中心となり，IoT 製品に対するセキュリティ適合性評価制度[]の構築を進めている．

一方，AI についても，欧州 AI Act[4]などの規制が始まっている．これは，AI 利用における倫理観やプライバシーへの配慮等が中心であるが，これらを実現する上で，システムのサイバーセキュリティが維持されていることは前提条件となっている．加えて，AI には，敵対的攻撃，ポイズニング，モデル盗取等の，AI 特有のセキュリティリスクがあり，これらについても考慮が必要である．

改めて法規を見ると，CRA では，デジタル要素として AI を利用する場合，そのセキュリティについては AI Act の内容を参照することが要求されている．AI Act では，第 15 章において，High-risk AI Systems の Accuracy, robustness および cybersecurity が要求されており，AI 特有のセキュリティとして，data poisoning, model poisoning, adversarial examples, model flaws が挙げられている．

自動車に至っては，欧州における自動車型式指定制度((EU)2019/2144)の下で安全性の確認が行われるが，AI Act により，AI が Safety に関係する場合，例えば自動運転車の自動運転機能において AI が用いられる場合は，その AI システムは High-risk AI Systems とみなされ，AI Act の要件が課される，という変更が行われている．また，UNECE WP.29 の GRVA 分科会では，自動車における AI の利用に関するガイダンスを作成中で，具体的には AI システムのソフトウェア更新，および学習に用いるデータの保護に関する推奨事項が記載されている．

これまで述べた通り，法規上は AI システムのサイバーセキュリティ，ならびに AI システム特有のセキュリティについて配慮することが要求されているが，具体的にどうするかまでは示されていないのが現状である．

2. サイバーセキュリティケースの作成

2.1 UN-R155 における Cybersecurity Case

UN-R155 では，7.3 節において，次のような要件が記載されている．

- CSMS 適合証明書を持つこと
- サプライチェーン管理に関する説明
- 自動車のリスクアセスメントの結果の説明

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

- リスクアセスメントの結果に基づく対策の説明
- 3rd パーティ製ソフトウェア実行環境の保護
- インシデント対応手順
- 十分なテストを実施したことの説明
- 暗号モジュールのコンセンサス標準への準拠

CSMS は，リスクアセスメントとインシデント対応を含む，組織のサイバーセキュリティ管理能力のことで，CSMS で規定された手続きに従い，リスクアセスメントやインシデント対応を実施する必要がある．車両開発におけるリスクアセスメントについては，次に挙げる ISO/SAE 2143 がスタンダードとなっている．

2.2 ISO/SAE 21434

ISO/SAE 21434 は，車両の E/E システムを対象としたサイバーセキュリティエンジニアリングの規格である．機能安全の規格である ISO 26262 をベースに検討されており，リスクアセスメントについては ISO 26262 の手法を参考としている．例えば，E/E システムに対するセキュリティ上の脅威と，脅威が顕在化した時の道路利用者(車両ユーザ等)への損害を区別し，脅威の発生率と損害を掛け合わせることでリスク値を算定する手法を取る．脅威分析については，ISO/IEC 27001 を参考に，E/E システム上の資産に対する CIA の侵害を軸に分析を行う例が示されている．

2.3 AI を利用する車両システムのリスクアセスメント

ISO/SAE 21434 は，対象となるシステムの種類や構成に依存しない規格であるため，自動運転機能を実現する E/E システムにおいても本プロセスを適用する．ここで課題となるのが，AI 特有の脆弱性に関するリスク分析のやり方である．AI 特有の脆弱性は，先に述べた AI Act に記載されている項目，もしくは，ISO/IEC TR 24028 に記載されている以下の 4 項目が参考となる．

- Data Poisoning
- Adversarial Attacks
- Model Stealing
- Hardware-focused threats to confidentiality and integrity

このうち，Hardware-focused threats については，情報セキュリティのリスクアセスメントの範疇である．

竹内らは，AI を利用する E/E システムとして ADAS を取り上げ，ADAS に対するリスク評価を行っている[5]．ADAS システムをモデル化し，そのうち AI を利用する認識部，フュージョン部，制御値計算部に対する AI 関連の脅威を検討し，CIA をセキュリティ特性を割り当てている．溝口らも，AI 特有の脅威を ISO/SAE 21434 のプロセスで扱う際に，CIA を割り当てる，あるいは CIA 以外の軸でプロセス・手順を構築する必要があることを述べている[6]．

攻撃実現可能性(Attack Feasibility)については，竹内らの主張の通り，認識部へ敵対的サンプルの入力が，セーフティへの影響および攻撃のしやすさの観点でリスク値が高くなる．また，敵対的サンプルの生成のために，認識部で用いられるモデル盗取/推定も脅威となる．

2.4 リスク対応

敵対的攻撃に対する対策は，竹内らが述べている通り，入力データの無毒化や，モデルのロバスト化といった予防的対策だけでなく，攻撃が成功した際に機能安全の考え方で検知・対応するやり方が考えられる．このような，AI システムの機能安全に関する文書が，ISO/IEC TR 5469 や ISO/PAS 8800 である．ISO/PAS 8800 は，ISO/IEC TR 5469 をベースに自動車向けに拡張した規格であるが，2024 年 8 月時点で公開されていない．

ISO/IEC TR 5469 は，non-AI システムの機能安全規格である IEC 61508 を，AI 利用システムに適用する上での課題についてまとめている．その中で，Adversarial Attacks についても言及されており，Adversarial Attacks のリスクに対応することが AI 利用システムの機能安全に寄与するという主張になっている．特に，Adversarial Examples に含まれる摂動を検知する手法として，MagNet や Defense-GAN と言った手法が例示されている．Adversarial Attacks を考慮した安全関連系のアーキテクチャとして，データの入力を，主機能となる AI/ML コンポーネントだけでなく，Supervisory components に入力して，システムの挙動に制限を設ける方式が記載されている．また，AI の説明可能性についても，機能安全を語るうえで重要な要素として挙げられている．

3. おわりに

AI を利用するシステムのセキュリティ論証は，機能安全を含めた幅広い対応が必要であり，今後も標準化動向を踏まえた調査が必要である．

参考文献

[1] UN-R155, https://unece.org/transport/documents/2021/03/standards/un-regulation-no-155-cyber-security-and-cyber-security

[2] 道路運送車両の保安基準, https://www.mlit.go.jp/jidosha/jidosha_fr7_000007.html

[3] EU Cyber Resilience Act, https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

[4] AI Act , https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[5] 竹内ら, “車載システムにおける AI に対するセキュリティリスク評価“., SCIS 2023.

[6] 溝口ら, “自動車の機能として AI/ML を利用する場合のセキュリティ論証について”, JSAI 全国大会 2023.