AIセキュリティに関する包括的な研究開発

研究代表者　森　達哉
（基幹理工学部 情報通信学科　教授）

1. 研究課題

人工知能（AI）の社会実装が急速に進む一方、深層ニューラルネットワーク、大規模言語モデル（LLM）、マルチモーダル生成モデルといった AI システムには、従来のサイバー攻撃では捉えきれない新たなセキュリティ・プライバシー脅威が顕在化している。加えて、LLM を悪用した高度なサイバー攻撃や、AI を活用したサイバー防御技術の出現により、攻守双方が AI 化する構造転換が進行している。このような状況下では、攻撃・防御・運用の全側面から AI セキュリティを体系的に捉え、実運用可能な防御技術を確立することが急務である。

本研究課題では、「AI とセキュリティ」を包括的な視点で捉え、以下の三領域に関する研究開発を推進する。(1) AI システムに対する脅威の体系的評価：自動運転等のフィジカル AI システムに対する物理的敵対的攻撃、マルチモーダル生成モデルに対するデータ汚染攻撃、医療 AI に対する敵対的信号攻撃、機械学習ベースマルウェア検出器に対する回避攻撃を対象とする。(2) AI を活用したセキュリティ防御技術の高度化：LLM エージェントによる悪性パッケージ解析、軽量 LLM によるフィッシング検知、LLM ツール呼び出し監査による悪性 MCP サーバ検知などを扱う。(3) AI のライフサイクル全体に関わるセキュリティ・プライバシー：AI 開発現場（GitHub Actions 等）のセキュリティ実態調査、LLM の安全性アライメント強化、プライバシーポリシー理解のための LLM 応用を対象とする。これら三領域の成果を相互に関連付け、AI セキュリティという新領域の学術基盤を確立することを目指す。

2. 主な研究成果

2.1　フィジカル AI システムに対する敵対的攻撃の体系的評価

自動運転に用いられる LiDAR センサの点群前処理フィルタに着目し、物理的な霧粒子を用いて AI 物体認識を誤認させる新たな攻撃 Adversarial Fog を提案した。本成果は ACM ASIACCS 2025 に採択され Best Paper Award を受賞した。さらに、物理影を悪用して LiDAR 物体検出を欺く攻撃 Shadow Hack を USENIX Security 2025 に発表し、攻撃の成立条件と防御手法を体系的に評価した。加えて、自動運転の経路計画 AI に対する敵対的軌跡攻撃 AVATAR、信号機認識モデルに対する Traffic-Light Hiding Attack、Occupancy Prediction に対する攻撃、HD マップ改ざん攻撃、V2X 連携下の攻撃など、自動運転 AI を構成する各モジュールに対する多面的な脅威を明らかにした。これらの研究は、AI を用いた自律走行システムの安全性保証に不可欠な脅威モデルを提供するものである。

2.2　マルチモーダル生成 AI に対するデータ汚染攻撃

Text-to-Image（T2I）生成モデルに対し、学習データ汚染を用いたバックドア攻撃が、英語以外の 10 言語において英語と比べ顕著に高い成功率を示すことを実証した。多言語環境が AI セキュリティ上の新たな脆弱性次元となる点は従来十分に議論されておらず、本研究はこの領域に基礎的知見をもたらすものである。本成果は ICPR 2026 に採択され、NDSS 2026 においてポスター発表を

行った。これにより、言語的多様性を前提としたAIモデルの防御設計原則を提示した。

2.3 AIを活用したセキュリティ業務の自動化

LLMエージェントによって悪性PyPIパッケージを自律解析するフレームワークCHASEを構築し、AIware 2025（ASE 2025併催）に採択された。また、軽量LLMをタスク最適化してフィッシング攻撃のブランドなりすましを検知するBrandSpotterをIEICE Trans. Information and Systemsに発表した。さらに、LLMエージェントのツール呼び出しログとOS監査ログを対応付けて悪性MCPサーバを検知する手法や、AIエージェントと機械学習を組み合わせたハイブリッド型フィッシングサイト検出システムも提案し、LLMの推論・計画能力をセキュリティ運用へ応用する先駆的な取り組みを体系的に展開した。

2.4 機械学習マルウェア検出器のロバスト性評価

Androidマルウェア検出器に対する低クエリ予算の転移型敵対的攻撃Noise-Augmented Transferability（IEEE Access）と、強化学習を用いてAPIパラメータを注入しアテンション型動的検出器を回避する手法（ACM SAC 2026）を提案した。さらに、セマンティクス保持型のマルウェア変種を生成し検出器のロバスト性を評価する研究（CSS 2025優秀論文賞）も並行して実施し、機械学習ベース検出器の安全性評価枠組みを精緻化した。

2.5 医療AIおよびAI時代の開発環境のセキュリティ

心電図に基づく自動不整脈診断AIに対し、敵対的音響信号によって偽の不整脈を誘発する攻撃Adversarial Beatsを実証し、ACM Trans. Cyber-Physical Systemsに掲載された。また、GitHub Actionsを題材とするAI時代の開発ライフサイクルのセキュリティ実態を混合研究法で調査する成果がNDSS 2026に採択され、ソフトウェア開発現場のセキュリティ課題を明らかにした。

3. 研究業績

3.1 学術論文（査読ありジャーナル、国際会議プロシーディングス）

1. Taiga Ono, Takeshi Sugawara, Jun Sakuma, Tatsuya Mori, “Adversarial Beats: Feasibility Study of Spoofed Arrhythmia,” ACM Trans. on Cyber-Physical Systems, Vol. 10, No. 1, Article 12, pp. 1–30 (2026).
2. Ryusei Watanabe, Mamoru Mimura, Takahiro Matsuki, Tatsuya Mori, “BrandSpotter: A Framework for Identifying Targeted Brand Names in Phishing Attacks Using Task-Optimized Lightweight LLMs,” IEICE Trans. on Information and Systems (2025).
3. Junji Wu, Tomohiro Morikawa, Tatsuya Mori, “Noise-Augmented Transferability: A Low-Query-Budget Transfer Attack on Android Malware Detectors,” IEEE Access, Vol. 13 (2025).
4. Yuna Tanaka, Kazuki Nomoto, Ryunosuke Kobayashi, Go Tsuruoka, Tatsuya Mori, “Adversarial Fog: Exploiting the Vulnerabilities of LiDAR Point Cloud Preprocessing Filters,” in Proc. ACM ASIACCS, August 2025. **[Best Paper Award]**
5. Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, Tatsuya Mori, “Invisible but Detected: Physical Adversarial Shadow Attack and Defense on LiDAR Object Detection,” in Proc. USENIX Security Symposium, August 2025.
6. Takaaki Toda, Tatsuya Mori, “CHASE: LLM Agents for Dissecting Malicious PyPI Packages,” in Proc. AIware 2025 (ASE Workshop), October 2025.
7. Ryohei Kakebayashi, Tatsuya Mori, “Cross-Lingual Vulnerabilities of Text-to-Image Models: Evaluating Data Poisoning Attacks Across Ten Languages,” in Proc. ICPR 2026,

September 2026 (to appear).

8. Junji Wu, Tomohiro Morikawa, Tatsuya Mori, “Evading Attention-based Dynamic Malware Detectors: An RL-Guided Method for Adversarial API Parameter Injection,” in Proc. ACM SAC, March 2026.
9. Yusuke Kubo, Fumihiro Kanei, Mitsuaki Akiyama, Takuro Wakai, Tatsuya Mori, “Action Required: A Mixed-Methods Study of Security Practices in GitHub Actions,” in Proc. NDSS 2026, February 2026.
10. Jiadong Liu, Tatsuya Mori, “AVATAR: Adversarial Vehicle Trajectory Attack Targeting Autonomous Driving Planner,” in Proc. Automotive Cyber Security Workshop (ACSW), June 2025.

国内学会発表

1. 戸田宇亮, 若井琢郎, 森達哉,「悪性パッケージ検出を目的とした Supervisor 型 LLM エージェントによる Python パッケージの自律的な段階的解析」, コンピュータセキュリティシンポジウム（CSS）2025, 2025 年 10 月. **[CSS2025 優秀論文賞]**
2. 鶴岡豪, Qi Alfred Chen, 野本一輝, 小林竜之輔, 田中優奈, 森達哉,「ヘッドライト光の反射を悪用した標識認識への攻撃：商用車両を用いた影響評価と対策の提案」, CSS 2025, 2025 年 10 月. **[CSS2025 優秀論文賞]**
3. 渡邉龍星, 富永大智, 犬塚祥, 森達哉,「防御技術強化を狙いとしたセマンティクス保持型マルウェア変種生成・拡張手法」, CSS 2025, 2025 年 10 月. **[CSS2025 優秀論文賞]**
4. 掛林諒平, 森達哉,「Text-to-Image モデルに対するデータ汚染攻撃の多言語環境における脆弱性要因の分析と緩和手法の提案」, 信学技報 ICSS2025-118, 2026 年 3 月.

3.2　総説・著書

該当なし

3.3　招待講演

1. Tatsuya Mori, “Security and Safety in Autonomous Driving: Present and Future Challenges,” CyberC3 Intelligent Vehicle Labs, Shanghai Jiao Tong University, October 2025.
2. Tatsuya Mori, “Security and Privacy of Connected Vehicles: Gaps in Perception among AI, Humans, and Developers,” IEEE/RSJ IROS 2025 Workshop on Are You Happy with AV?, October 2025.
3. Tatsuya Mori, “AI Security in Practice: From Securing Autonomous Systems to AI Agents for Malware Analysis,” AI Security and Privacy Team Research Workshop, September 2025.
4. 森達哉,「新興技術時代のオフェンシブセキュリティ研究と人材育成の課題」, 第 22 回情報セキュリティ文化賞記念講演会（日経メッセ 2026）, 2026 年 3 月.
5. 森達哉,「AI とセキュリティの将来課題」, NICT サイバーセキュリティシンポジウム 2026, 2026 年 2 月.
6. 森達哉,「AI とセキュリティ」, JNSA 設立 25 周年記念イベント講演会, 2026 年 2 月.
7. 森達哉,「フィジカル AI セキュリティ：オフェンシブセキュリティ研究の現在地と展望」, SCAIS 2026, 2026 年 1 月.
8. 森達哉,「人工知能の観点からの次世代セキュリティ」, 第 12 回 ASF 次世代セキュリティシ

ンポジウム, 2025 年 12 月.

9. 森達哉,「身近なサイバー攻撃の現状と最新 AI 技術を活用した対策技術への期待」, サイバーセキュリティ経営戦略セミナー（兵庫県立大学）, 2025 年 12 月.

3.4 受賞・表彰

1. 令和 7 年度 科学技術分野の文部科学大臣表彰, 文部科学省, 森達哉, 2025 年 4 月.
2. 第 22 回 情報セキュリティ文化賞, 情報セキュリティ大学院大学, 森達哉, 2026 年 1 月.
3. ACM ASIACCS 2025 Best Paper Award, “Adversarial Fog: Exploiting the Vulnerabilities of LiDAR Point Cloud Preprocessing Filters,” Y. Tanaka, K. Nomoto, R. Kobayashi, G. Tsuruoka, T. Mori, ACM, August 2025.
4. USENIX VehicleSec 2025 Best Demo Award, G. Tsuruoka et al., USENIX, August 2025.
5. CSS 2025 優秀論文賞（3 件）: 鶴岡豪ほか, 戸田宇亮ほか, 渡邉龍星ほか, 情報処理学会, 2025 年 10 月.
6. CSS 2025 学生論文賞（2 件）: 海老根祐雅ほか, Lachlan Moore ほか, 情報処理学会, 2025 年 10 月.
7. CSS 2025 奨励賞（2 件）: 久保佑介ほか, 迫本澪ほか, 情報処理学会, 2025 年 10 月.

3.5 学会および社会的活動

- 総務省　サイバーセキュリティタスクフォース　委員（2025 年 9 月～）
- 情報処理学会　コンピュータセキュリティ研究会（CSEC）　主査（2025 年 4 月～）
- IWSEC 2026　General Co-chair（2025 年 12 月～）
- 科学技術振興機構（JST）研究開発戦略センター　分野別委員会　委員（第 1 通信分野）
- JST　CREST「人工知能と情報社会」領域　領域アドバイザ（CRONOS）
- JST　さきがけ「社会変革に向けた ICT 基盤強化」　領域アドバイザ
- 電子情報通信学会　情報通信システムセキュリティ研究専門委員会　研究専門委員
- 電子情報通信学会　情報セキュリティ研究会　研究専門委員
- 理化学研究所　革新知能統合研究センター（AIP）　人工知能セキュリティ・プライバシーチーム　客員研究員
- 情報通信研究機構（NICT）　サイバーセキュリティ研究所　招へい専門員

4. 研究活動の課題と展望

2026 年度も引き続き、AI セキュリティの三領域（攻撃・防御・運用）に関する包括的研究を推進する。(1) LLM 基盤自律エージェントのセキュリティについては、ツール呼び出し経路の詳細監査によるサプライチェーン攻撃検知、悪性 MCP サーバに対する防御、LLM の安全性アライメント強化等について、理論と実装の両面から研究を深化させる。(2) フィジカル AI システム（自動運転・医療 AI・スマートインフラ等）については、V2X 通信を含むシステムレベルの脆弱性評価、物理的な光学効果や敵対的背景を用いた攻撃の実環境下評価、及び運用環境を想定した体系的防御設計を推進する。(3) 生成 AI の安全性については、多言語環境におけるデータ汚染耐性、マルチモーダル入力に対する敵対的堅牢性、生成物の識別可能性や倫理的課題への対応に取り組む。(4) AI が組み込まれたソフトウェア開発ライフサイクル（GitHub Actions、CI/CD パイプライン等）のセキュリティ実態調査と、機械学習・LLM 併用型の悪性ソフトウェア検知システムの大規模評価も継続して進める。これらを通じ、AI セキュリティの理論・計測・防御・運用を横断する学術基盤を構築し、安全で信頼できる AI の社会実装に寄与する。