国際会議 IJCAI2024/AI-Safety 参加発表報告

櫻井　幸一 [1,*]

概要：2024 年 8 月に韓国は済州島で開催の IJCAI ワークショップ/AI-Safety で発表した際の会議報告を行う。

キーワード：人工知能, 安全、安心、セキュリティ、IJCAI

A report on participation and presentation in IJCAI2024/AI-Safety

Kouichi SAKURAI [1,*]

Abstract: This reports the author's attending to IJCAI Workshop/AI-Safety in Jeju, Korea in August 2024, in which the author give a presentation.

Keywords: AI, Safety, Trust, Security, IJCAI

1. はじめに

IJCAI 本会議の参加報告は、人工知能学会誌に掲載されている。かつ参加速報も発表者自身のブログで多数公開さている。しかし、AI-safety に関しては、まだ４回目と歴史が浅いためか、国内研究者による報告が皆無のため、ここに情報提供する。

今回の対面参加は、採択された論文発表(2024 年 3 月 CSEC 千葉工大での下名の研究発表の英語版/Arxiv)が主目的であった。また、日本からは、筑波大の参加発表もあった。

2. セキュリティ AI 系ワークショップ群

セキュリティ AI 系のワークショップは、他の主要 AI 系会議にも併設されている：

WAISE -Organized at SAFECOMP: 2024 年/7th は 9 月にイタリアにて

SafeAI-Organized at AAAI: 2023 年は 2 月にワシントン DC にて開催済み。2024 は SafeAI ではなく、特別トラック Safe, Robust and Responsible Artificial Intelligence (SRRAI)が企画されている。

SAIAD -Organized at ECCV : Safe Artificial Intelligence for All Domain (2022 年までの前身は Safe Artificial Intelligence for Automated Driving) CVPR の併設として、2024 年/6th は 6 月にシアトルで開催されている。

3. スコープと話題

AISafety の HP/CFP に記載のスコープと扱う話題は以下の通りである[著者の抄訳/抜粋]。
人工知能(AI)のリスクに対する懸念は高まっている。インテリジェントシステムの意思決定や制御ループから人間が徐々に排除されるにつれ、安全性の重要性はますます高まっている。実際、生成 AI、大規模言語モデル、機械学習などの基盤となる AI 技術やアルゴリズムは、有害または虚偽の反応で人間の価値を損なう可能性がある。利用される AI アルゴリズムが物理世界や人間自身と対話し、より高い自律レベルで活動しているシステムに対しては、既存の安全性工学の原理に基づく技術基盤と仮定では、不十分な時代である。また、現在の AI システムがもたらす安全性の課題と、汎用人工知能を含むより高性能な AI システムに焦点を当てたより将来とを見据えた研究との関連性も考慮すべきである。

このワークショップでは、人工知能の安全性に関する新しいアイデアを探求し、特に次の質問群に対処することに焦点を当てる。

- **信頼できる AI** ソフトウェアアーキテクチャをどのように設計できるのか?
- AI 基盤システムをより倫理的に整合させるために、システムエンジニアリングにおいて制限された道徳を特定し、使用する必要はあるか?
- AI と ML の安全性を確保するための既存のアプローチの状況とギャップは何か?

1 九州大学 Kyushu University
* sakurai@inf.kyushu-u.ac.jp

- 自動化された意思決定システムで安全な人間と機械の相互作用を開発するには、どのような安全性工学的考慮事項が必要か?
- AI の安全性に関する業界からの考慮事項と経験は何か?
- AI システムをその潜在的なリスクと脆弱性に応じてどのように特徴付け、評価できるか?
- AI の安全性に関する確固たる技術的ビジョンとパラダイムシフトの記事をどのように開発できますか?
- 能力と一般性の指標は、システムのリスクレベルにどのように影響し、パフォーマンスとのトレードオフをどのように見つけることができるか?
- AI システムの特徴は、例えば、倫理、説明可能性、透明性、説明責任など、その安全性にどのように関連し、または貢献しているか?
- AI の安全性を評価するには?
- 生成 AI・大規模言語モデル・機械学習を保護する方法は何か?

このワークショップの主な関心は、AI や安全性工学などの複数の分野を倫理的および法的問題とともに総合的に考慮し、信頼できるインテリジェントな自律型機械を構築する必要があるシステムエンジニアリングの新しい視点を見ることである。

AISafety では、以下を含むがこれらに限定されない AI の安全性のあらゆる側面をカバーする理論的、実験的、およびポジションペーパーを募集する。

- AI に基づくシステムアーキテクチャにおける安全性
- 継続的な V&V と AI の安全特性の予測可能性
- ランタイム監視と AI 安全性の(自己)適応
- AI に基づくシステムの説明責任
- 説明可能な AI と解釈可能な AI
- AI の安全リスクの検出と軽減
- AI に基づくシステムにおける悪影響の回避
- 監視の役割と有効性:相関性と中断可能性
- 価値の喪失と壊滅的な忘却の問題
- 自信、自尊心、そして分布シフトの問題
- 汎用 AI システムの安全性と一般性の役割
- 報酬のハッキングとトレーニングの破損
- 自己説明、自己批判、透明性問題
- ヒューマンマシンインタラクションの安全性
- AI に基づくシステムの規制:安全基準と認証
- ヒューマンイン/オン/アウトオブザループとスケーラブルな監視の問題
- ミックスド・イニシアティブ・コントロール・フレームワーク
- AI の安全性評価プラットフォーム
- AI 安全教育と意識向上
- 産業、健康、自動車、航空宇宙、ロボット工学などの分野を含む、AI に基づくセーフティクリティカルなシステムでの経験
- 堅牢性、安定性、信頼性、制御性などの安全関連特性の検証に焦点を当てた、最近の AI 規制や規格に準拠するためのアプローチ

4. 予稿集

本ワークショップの予稿集は CEUR-WS(http://ceur-ws.org/)に掲載される[2024 年 8 月 22 日の時点で 2024 年版はまだ追加されていない]。CEUR-WS は、論文が出版されると削除できないという意味で「アーカイブ」であるということ。著者は、CC BY 4.0 に従って論文の著作権を保持する。つまり、CEUR-WS は Arxiv に似ている。いずれにせよ、受理された論文の著者は、オプトアウトして、自分の論文を議事録に含めないことを決定することができるということである。

IJCAI に受理されなかった論文の投稿も歓迎すると明示ある。

なお、発表スライドは、事前に会議長に送る指示が出され、それに従った。結果として、他の一般論文スライドも、会議までには共有できた。

また、基調講演の内２件のスライドも、直接講演者にお願いし、送ってもらうこともできた。

5. 講演概要

一般講演は、発表時間 10 分と短めである。質疑は、各セッション最後のパネル討論という形式で、講演者全員が前に出て、会場からの失認へ回答するという形式であった。
当初は、2 日間の枠をとってのワークショップであったが、結果的には 1 日のみ終日の開催となった。
また、基調講演 4 つのうち、2 件は online, 1 件は実現しなかった。

以下では、基調講演と各講演の概要を解説する。
基調講演が三つ（プログラムではもう一件が予定されていた)。

基調講演/その１：The Evolvement of AI/ML Aviation Regulations and Illustration of Some Practical Aspects through an End-to-End Certification Case Study Dr. Konstantin Dmitriev (Technical University of Munich)
人工知能、特に機械学習 技術の急速な進歩により、航空機システムへの AI/ML の安全な統合を確保するための航空規制フレームワークの進化が必要になってきた。
近年では、航空当局、標準化団体、業界関係者が
協力し、航空機搭載 ML 応用システムを認証するための
規制枠組みを確立してきた。
本講演では、航空分野における AI 規制の動的な状況を探り、すでに確立されている重要な要素と進行中の取り組みに焦点を当てる。これらの新たな規制の実際の応用を説明

するために、空港標識の視覚的検出と分類に深層学習を採用した ML ベースの航空機緊急ブレーキシステム (AEBS) の事例を紹介する。
講演では、ML 開発保証プロセス、アーキテクチャ上の緩和戦略、データ管理の実践、ML モデルの意図しない動作、安全性評価に対する ML の性能制限と変動性の影響など、ML 固有の側面に焦点を当てる。 AEBS 事例は、EUROCAE WG-114 / SAE G-34 共同作業グループ「航空における人工知能」内で、議論を促進し、さまざまな規制トピックの収束を促進するための例として活用されている。

基調講演その 2:
Journey and Findings of the Research Program Confiance.ai Loïc Cantat(IRT SystemX)
フランス政府のイノベーション評議会によって立ち上げられた「人工知能に基づくシステムの信頼性の確保、認定、強化」というグラン・デフィの技術的柱であるプログラム Confiance.ai の歩みと成果を紹介する。データ管理 (データ品質、データ生成など) と堅牢性の特性に焦点を当て、Confiance.ai の活用事例を取り上げる。
[https://www.confiance.ai]

基調講演その３ :
Ability-Centered AI And Policy (Transatlantic Safety Dialogue And Designated Groups) Yonah Welker(EU Commission projects, Yonah.org)
AI 法の既存の問題、顔の非対称性、多様なジェスチャー、コミュニケーション スタイル、行動、行動パターンを持つ個人のレンズを通して、許容できない高リスクシステムの課題を取り上げる。特に、障害のある人々、認知障害および感覚障害、自閉症スペクトラム障害。また、アルゴリズム、警察と市のシステムのカテゴリー、提案されている行動と基準 (支援技術と障害者中心の AI システムの促進に関する 6 つ、安全性と誤用の防止に関する 8 つ)、監査とコンプライアンスの枠組みなど、誤用とサイロ化に対処する統計も取り上げています。 (*EU 障害者フォーラムを含む EU 150 機関が署名した公開書簡に続く) AI システムが特定の出身地や肌の色を持つ人々を差別するのと同様に、コンピューター ビジョン、顔認識、音声認識、雇用または医療プラットフォームなどのシステムも、障害のある人を差別する。顔の違いや非対称、異なるジェスチャー、ジェスチャー、言語障害、または異なるコミュニケーションスタイルは、不正確な識別や差別につながる可能性がある。実際の報告事例を通しての課題は、関係するグループの公平性、透明性、説明可能性、人間中心性、プライバシーとセキュリティを確保する「障害中心」または「神経多様性中心」の研究、開発、監査の枠組みの必要性につながる。

6. 一般講演と論文賞

一般講演は、４セッション８件の発表であった。

ReLess: A Framework for Assessing Safety in Deep Learning Systems
Nan Jia[1], Anita Raja[1,2], Raffi Khatchadourian[1,2]
[1]The Graduate Center, CUNY [2]Hunter College, CUNY

Enhancing Autonomous Vehicle Safety through N-version Machine Learning Systems
Qiang Wen[1], Júlio Mendonça[2], Fumio Machida[1], Marcus Völp[2]
[1]Department of Computer Science, University of Tsukuba
[2] Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg

Hyper-parameter Tuning for Adversarially Robust Models
Pedro Mendes[1,2], Paolo Romano[1], David Garlan[2]
[1]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
[2] Software and Societal Systems Department, Carnegie Mellon University

Low-Latency Privacy-Preserving Deep Learning Design via Secure MPC
Ke Lin, Yasir Glani, Ping Luo (Tsinghua University)

Detecting out-of-distribution text using topological featuresof transformer-based language models
Andres Pollano(Melbourne University) Anupam Chaudhuri(Deakin University) Anj Simmons (Hashtag.AI)

The Impact of Prompts on Zero-Shot Detection of AI-Generated Text
K.Kaito[1], Y. Gu and K. SAKURAI (Kyushu University)
[1] Now with SkyDisc.

Global Clipper: Enhancing Safety and Reliability of Transformer-based Object Detection Models
Qutub Syed[1,3] , Michael Paulitsch[1] , Karthik Pattabiraman[2] , Korbinian Hagn[1] , Fabian Oboril[1] , Cornelius Buerkle[1], Kay-Ulrich Scholl[1] , Gereon Hinz[3] and Alois Knoll[3]
[1]Intel Labs, Munich, Germany
[2] University of British Columbia, Vancouver, Canada
[3] Technical University of Munich, Munich, Germany

Neural Vicinal Risk Minimization: Noise-robust Distillation for Noisy Labels
Hyounguk Shon[1], Seunghee Koh[1], Yunho Jeon[2], Junmo Kim[1]
[1]Korea Advanced Institute of Science and Technology
[2]Hanbat National University

最優秀論文賞は、事前に二の候補
ReLess: A Framework for Assessing Safety in Deep Learning Systems (CUNY) と
Hyper-parameter Tuning for Adversarially Robust Models (CMU)
とが告知されていた。
会議当日のワークショップ長らの投票で後者[Arxiv でも公開]が受賞している。
この論文賞の研究では、ロバスト (敵対的に訓練された)

モデルのハイパーパラメータチューニング（HPT）の問題に焦点を当てている。このため、3 つの一般的なディープモデルに基づく広範な実験的研究を実施し、9 つの（離散化された）ハイパーパラメータ（HP）、2 つの忠実度次元、および 2 つの攻撃境界を、合計 19208 構成(50,000GPU 時間に相当）を調査している。
この研究を通じて、標準的な訓練と敵対的な訓練で使用する HP を独立に調整する必要があるため、敵対的な設定では HPT 問題の複雑さがさらに悪化することを示している。これを成功させる（つまり、両方のフェーズで異なる HP 設定を採用する）ことで、クリーン入力と敵対的入力の誤差をそれぞれ最大 80%と 43%削減できると主張している。また、ロバストモデルの HPT のコストを削減する新たな機会、具体的には、安価な敵対的訓練法を活用することで、よりロバストで高価な最先端の手法で達成可能な品質の、安価で相関性の高い推定値を得ることも提案している。この新しいアイデアを最近のマルチフィデリティオプティマイザ（taKG）と併用することで、HPT プロセスの効率が最大 2.1 倍まで向上することも示している。

7. IJCAI/チュートリアル/セキュリティ関係

IJCAI は、本会議は 4 日目からで、前半の 3 日間は、ワークショップ（WS）とチュートリアル（TT）31 件とワークショップ(WS)３6?件が並列で開催された。下名が参加発表した AIsafety は、2 日終日（am09:00--17:30）を使ってのプログラムであった。
関連して、3 日目に参加した二つのチュートリアルを報告する。

T26: Toward Mitigating Misinformation and Social Media Manipulation in Foundation Model Era
ソーシャルメディア上で世論を操作するための誤報の一貫した悪用は、様々な領域でますます顕著になってきている。この脅威は、大規模言語モデル（LLM）の発達により、操作者がより効率的に説得力の高い欺瞞的なコンテンツを生成できるようになったため、深刻さを増している。さらに、ChatGPT のような LLM と統合されたチャットボットの最近の進歩は、人間のような対話型ソーシャルボットの作成を可能にし、人間のユーザーとソーシャルメディアプラットフォームのソーシャルボット検出システムの両方に重要な課題を投げかけている。このチュートリアルでは、(1)社会的操作や誤情報発信行動をモデル化するための機械学習、(2)LLM によって生成される誤情報、(3)LLM に基づく誤情報検出など、この目標に役立つ先進的な機械学習研究を紹介する。さらに、今後の方向性についても述べる。
研究スライドも公開されている
[https://www2024.thewebconf.org/docs/tutorial-slides/toward-mitigating-misinformation.pdf]

T27: A Copyright War: Authentication for Large Language Models
LLM の普及により、(1) 教育および学術における盗用、(2) 誤情報の流布、(3) モデル抽出攻撃など、デジタル コンテンツと知的財産の完全性を脅かす誤用に関する懸念が高まっています。このチュートリアルでは、これらの問題の核心に迫り、LLM によって生成されたコンテンツの著者を認証するように設計された最先端技術の概要を紹介する。これらの対策を検討することで、参加者は自分の作業を保護し、これらの強力なモデルを責任を持って使用するための貴重な洞察を得ることがでる。

参加していないが、本会議での， Safety， Security, Adversarial attacks, Privacy, Trust 関係のセッションを以下に抜粋する。 ほとんどは Arxiv に preprint が公開され、その一部は、2023 年の主要会議でポスター発表されている。
ETF: AI Ethics, Trust, Fairness (1/2), ETF: Safety and robustness
MTA: Security and privacy, ETF: AI Ethics, Trust, Fairness (2/2), ETF: Trustworthy AI
抜粋した関連発表を付録に記す。

8. おわりに

チュートリアル T27 の後半は、下名の発表研究でも取り上げた LLM のプロンプトへの影響を理論的に解析する研究成果が紹介された。当初は発表主体の参加であったが、このチュートリアルで最新の関連研究の情報を得ることができたことが、結果的には、最大の成果物となった。

また、この T27 講演は、対面一名と online 一名で行われたが、USC(南カルフオルニア大)博士課程学生のグループであったことも目を引いた。

本シンポジム CSS や、CSEC 研究会で発表されている AI セキュリティやプライバシーに関する研究論文も、積極的に英語化され、ここで紹介した AISafety をはじめ、一線の国際会議へ投稿されることを期待する。

次の IJCAI2025 は、モントリオール（カナダ）での開催予定である。

付録 IJCAI 主会議での関係セッションと発表論文

ETF: AI Ethics, Trust, Fairness (1/2)

#J3 (12m) - Negative Human Rights as a Basis for Long-term AI Safety and Regulation (Abstract Reprint)
#SC8851 (12m) - Defending Against Backdoor Attacks by Layer-wise Feature Analysis (Extended Abstract)
#SV8084 (12m) - Robust Counterfactual Explanations in Machine Learning: A Survey
#3090 (6m) - Discriminative Feature Decoupling Enhancement for Speech Forgery Detection

#7338 (6m) - Speech-Forensics: Towards Comprehensive Synthetic Speech Dataset Establishment and Analysis#3265 (6m) - **ETF: Safety and robustness**

#SV7931 (12m) - Safety of Multimodal Large Language Models on Images and Text

#6053 (6m) - A General Black-box Adversarial Attack on Graph-based Fake News Detectors

485 (6m) - SwiftThief: Enhancing Query Efficiency of Model Stealing by Contrastive Learning

#1614 (6m) - Protecting Object Detection Models from Model Extraction Attack via Feature Space Coverage

#3666 (6m) - BADFSS: Backdoor Attacks on Federated Self-Supervised Learning

CV: Adversarial learning, adversarial attack and defense methods

#234 (12m) - Spear: Evaluate the Adversarial Robustness of Compressed Neural Models

#2331 (6m) - A Dataset and Model for Realistic License Plate Deblurring

#2053 (6m) - Imperio: Language-Guided Backdoor Attacks for Arbitrary Model Control

#2991 (6m) - Advancing Generalized Transfer Attack with Initialization Derived Bilevel Optimization and Dynamic Sequence Truncation

#1819 (6m) - Improving Adversarial Robustness via Feature Pattern Consistency Constraint

#418 (6m) - IntensPure: Attack Intensity-aware Secondary Domain Adaptive Diffusion for Adversarial Purification

August 07, 2024

MTA: Security and privacy

#3880 (12m) - Are Watermarks Bugs for Deepfake Detectors? Rethinking Proactive Forensics

#3057 (12m) - Delocate: Detection and Localization for Deepfake Videos with Randomly-Located Tampered Traces

#4663 (6m) - Shadow-Free Membership Inference Attacks: Recommender Systems Are More Vulnerable Than You Thought

#4454 (6m) - Privacy-Preserving UCB Decision Process Verification via zk-SNARKs

ETF: AI Ethics, Trust, Fairness (2/2)

#J12 (12m) - Performative Ethics From Within the Ivory Tower: How CSPractitioners Uphold Systems of Oppression (Abstract Reprint)

#4535 (6m) - Normative Testimony and Belief Functions: A Formal Theory of Norm Learning

#1194 (6m) - On the Effects of Fairness to Adversarial Vulnerability

#6438 (6m) - EAB-FL: Exacerbating Algorithmic Bias through Model Poisoning Attacks in Federated Learning

#2005 (6m) - When Fairness Meets Privacy: Exploring Privacy Threats in Fair Binary Classifiers via Membership Inference Attacks

ETF: Trustworthy AI

#3000 (6m) - PRASS: Probabilistic Risk-averse Robust Learning with Stochastic Search

#2006 (6m) - Ten Words Only Still Help: Improving Black-Box AI-Generated Text Detection via Proxy-Guided Efficient Re-Sampling

#2383 (6m) - Machine Unlearning via Null Space Calibration

#2617 (6m) - Pluggable Watermarking of Deepfake Models for Deepfake Detection

#5099 (6m) - BadFusion: 2D-Oriented Backdoor Attacks against 3D Object Detection