住友電工
Connect with Innovation

2nd GrandCanvas

AIマネジメントの国際標準化動向
～人工知能国際標準化シンポジウムから～

住友電気工業株式会社
サイバーセキュリティ研究開発室　三宅　和公
2024/07/31

SUMITOMO ELECTRIC GROUP

目次

1.自己紹介

住友電気工業株式会社　サイバーセキュリティ研究開発室
三宅　和公（みやけ　かずまさ）

現在
住友電工から産総研AIQM(機械学習品質マネジメント）に参加しAIセキュリティを担当。

経歴
- ブリッジ・ルータやスイッチなどインターネットのインフラ機器の組み込みソフトウェアを開発。(全レイヤ)
- GISシステム開発
  映像リアルタイム配信
  4K CATV関連システム・ASP設計・開発・保守・運用を経てセキュリティ研究開発へ。
  ASP開発ではセキュリティの最前線で24H365D、セキュリティ防衛実務に従事。
- 上記の傍ら、仲間を募って統計学やデータサイエンスを勉強。
  その社会実装の足掛かりとしてNEDO「実データで学ぶ人工知能講座」に2期連続参加したことをきっかけにAIセキュリティの研究開発へ。
  2020年からAIQMに参加。

2.はじめに

このトークでは、先日産総研が開催した人工知能国際標準化シンポジウムの内容を下敷きに米国のAIマネジメント・ガバナンスを中心に取り上げ、AIマネジメントのトレンドを紹介する。
（欧州委員会など米国以外の情報も簡潔に取り上げる。）

【ご注意】

- 発表の内容は、個人の調査・研究によるものであり、住友電工・産総研の組織としての見解ではない。
- 本書の記述のうち人工知能国際標準化シンポジウムの内容は青字で示す。

2.はじめに

人工知能標準化国際シンポジウム：Trustworthy AI 実現へ向けたAI標準化

International Symposium on AI Standards: towards trustworthy AI through standards

2024年7月3日 (水) ～ 5日(金)

3 July (Wed) – 5 July (Fri), 2024

日本科学未来館 7F イノベーションホール

Innovation hall, 7th floor, The National Museum of Emerging Science and Innovation

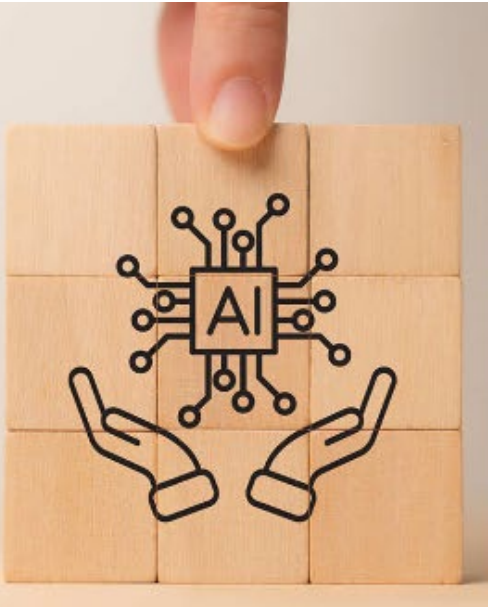
〒135-0064 東京都江東区青海2-3-6 日本科学未来館

2-3-6 Aomi, Koto-ku, Tokyo, 135-0064 Japan

(10時より前の入口は職員通用口をお使いください, entrance before 10:00 am: 35.619351N, 139.776093E)

主催：産総研人工知能標準化委員会（経済産業省の委託事業により開催）

Organizer: AI standardization committee, National Institute of Advanced Industrial Science and Technology (AIST), entrusted by the Ministry of Economy, Trade and Industry (METI)



https://www.digiarc.aist.go.jp/event/ai-standardization-symposium-2024/

3.概要

AIマネジメント・ガバナンスが脚光を浴びている。
不安定な社会情勢を背景にAIの利用についてリスクとなる事柄に世界各国が懸念を持っており、リスクが何かを捉え対策を導出する枠組みとしてAIマネジメント・ガバナンスの形になった。

【米国】
昨年秋頃に大統領令EO 14110が発出され、大統領令に沿う形でNIST AI RMF (NIST AI Risk Management Framework)の生成AI対応が2024年の春頃にリリースされ、更に大統領令に沿った発出文書のスケジュールを示し、政策と技術の両面で対策を打ち出しつつある。

【欧州委員会】
AI actで域内の利益保護を強化するとともにAIがもたらすリスクに対する不安からAI actに整合する国際標準としてISO/IEC 42001などの国際標準策定を進めた。

これらの情勢を紹介するイベントとして人工知能国際標準化シンポジウムが開催された。ここでは米国情勢を中心にAIマネジメントの動向を紹介し、米国以外の情報については簡潔に紹介する。

4.米国情勢

【政策面】

(1) 米国では昨年秋頃に大統領令EO 14110が発出され、AIのマネジメントに関する文書について「何をいつリリースするか」の発出予定を示した。

(2) 次いで米国は大統領令に沿う形でNIST AI RMFの生成AI対応を2024春頃にリリースし、生成AIのリスクへの対応を組み込んだ。

(3) その後、NISTは大統領令に沿った発出文書のスケジュールを示し、想定するリスクと対策の状況について情報を整理し発出し始めた。

4.米国情勢

NIST AI RMFの長であるElham Tabassi氏が大統領令EO 14110に関連する文書発行予定を紹介。
（下記でGAIはGenerative AIを指す。）

【大統領令EO 14110に関する文書発行予定】

2024年6月26日　合成コンテンツ認証に関する報告書をOMBおよびNSCに提出
2024年7月26日　①GAI向けAI RMFの公表
②GAIおよびデュアルユースモデルのためのセキュアソフトウェアフレームワークの発行
③AI能力を評価・監査するためガイダンス／ベンチマークを作成するイニシアチブを始動
④テスト環境の提供
⑤レッドチーミングガイドラインの発行
⑥産業界や関連する合成核酸配列のプライバイダとの連携を促進する。
⑦合成コンテンツ認証レポートの発行
⑧AI標準の推進と開発に関するグローバルな関与のための計画を公表する。
2024年10月29日 差分プライバシー保証保護の有効性に関するガイドラインの公表
2024年12月24日 合成コンテンツ認証ガイダンスの発行
2025年1月26日　標準化に関するグローバル・エンゲージメント計画に従って実施された優先的措置に関する報告書を大統領に提出

4.米国情勢

【技術面】
前のページで紹介した7/26発出のNISTの成果物
(1) NIST AI 800-1
“Managing Misuse Risk for Dual-Use Foundation Models”

- AIセキュリティを元に重大なリスクを紹介。AIセキュリティとしては回避攻撃・データ/モデルポイズニング・メンバーシップ推論などを紹介。
- リスクとしてディープフェイクやモーフィング（複数の人の写真を合成して合成元のどの人とも推論される写真を合成→旅券偽造など)、CBRN(Chemical, Biological, Radiological, Nuclear)への悪用、悪質なテキストの捏造など、AIセキュリティの脅威や脆弱性が安全保障上の危機に繋がることを訴える。
- 基盤モデルについても言及し、リスクの想定と対策(ガードレールの設置など)が現状で難しいことを述べる。
- Anthropicによる脅威のモデル化と脅威による被害の想定、悪用リスクの予測と管理計画・モデルの盗難リスク・悪用リスクの測定・配備後の悪用に関する情報　収集・悪用リスクに関する透明性の提供について述べる。

4.米国情勢

(2) NIST AI 100-2
“Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations”

- AIを「推論するAI」(PredAIと呼ぶ)と「生成するAI」(GenAIと呼ぶ)に分ける。
- PredAIの攻撃の分類を紹介：CIAそれぞれが侵害される攻撃を紹介。
- PredAIにおいて攻撃者の目的・目標の大分類を提示する。
- PredAIのおける攻撃として回避攻撃・データ/モデルポイズニング・プライバシーに関する攻撃を紹介する。いつ攻撃するか・何を侵害するか・攻撃者の能力や知識・モダリティ(画像・音声・テキスト・映像)
- GenAIの攻撃の分類を紹介：CIAそれぞれが侵害される攻撃を紹介。攻撃タイミングとして訓練時の攻撃と推論時の攻撃の分類も示す。
- GenAIの攻撃で攻撃者が活用しなければならない能力を示す。
- GenAIにおける攻撃を紹介：プロンプトインジェクション・ポイズニング攻撃など。

4.米国情勢

(3) NIST AI 600-1
“Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile”

- NIST AI RMFの生成AI対応。
- 組織がGenAIによってもたらされる固有のリスクを特定し、目標と優先順位に最も適したGenAIリスク管理のアクションを提案するガイダンスである。
- NIST AI RMFのユーザー向けの参考として有効なリソースとなることを目的とする。そのために12のリスクと200を超えるアクションのリストを定義する。
- 12のリスクには攻撃への参入障壁の低下、誤情報や偽情報、ヘイトスピーチ、その他の有害なコンテンツの作成、GenAIによる「幻覚」化(事実と異なる情報によって利用者を混乱に陥れる)が含まれる。
  →Appendixを参照。
- 開発者はリスクを軽減するためのアクションをAI RMF Core概念に対応する表で示す。

RMF Core：「MEASURE」で監視・分析・評価し見つけたAIリスクについて、「MAP」で関連するAIアクターと潜在リスクの洗い出しを行って原因予測して改善プロセスを構築し、「MANAGE」で対応リソースを割り当てて実施する。この一連の流れを「GOVERN」で方針・ルールに則り回す。

4.米国情勢

(4) NIST SP800 218A
“Secure Software Development Practices for Generative AI and Dual-Use Foundation Models”

- NIST SP800 218Aが規定するセキュアソフトウェア開発フレームワーク（SSDF）で定義されたセキュアソフトウェア(Secure Software Development Framework)の開発プラクティスとタスクを補強するもの。
- ソフトウェア開発ライフサイクル全体を通じて AI モデル開発に特化したプラクティス、タスク、推奨事項、考慮事項、注記、および参考文献を提供する。
- 大項目としては、
  - ・組織としての準備：ソフトウェア開発環境におけるセキュリティ要件定義など
  - ・プロテクトソフトウェア：開発するコードやデータを防衛する
  - ・安全性の高いソフトウェアの開発
  - ・脆弱性への対応

  から成る。

4.米国情勢

(5) NIST AI 100-5

“A Plan for Global Engagement on AI Standards”

国際標準策定の上で留意するべき事項を規定している。

- アクセスのしやすさ
- 科学的根拠
- ステークホルダが多様であること
- 人権への配慮

など。

4.米国情勢

【そのほか気になる情報】

- 共和党大会でのトランプの演説
- 7/26米国時間の発表内容

5.米国以外の情勢

【各国共通】

(1)AIマネジメント・AIガバナンスが最先端のトレンドとなった。世界規模であり、ISO/IEC 42001がリリースされた。

(2)AIマネジメントの認証に関する国際標準化が加速（ISO/IEC 17067ほか）

(3)ISO/IEC SC42による機能安全に関する活動が加速している。TR5469(AIの機能安全)、TS22440(ライフサイクル)において機械学習品質マネジメントガイドラインから貢献。

(4)AIのデータ(ISO/IEC 5259)活動進む。Accuracy, Completeness, Consistency, Credibility, Currentnessなどの特性を扱う。

(6)Human Oversight登場：AIの出力の統計的な性質(エラーとなるケース)を補完する監視の仕組みづくり

(7)Human Machine Teaming：人と機械学習の関係を5つのモデルに整理→要求仕様やKPI策定にヒントになるのでは。

5.米国以外の情勢

【欧州委員会】
- 2024年にAI act制定、2025年中に整合規格を開発。2026年にハイリスクAI対応。
- AI Trust Alliance創設。

【英国】
- AI Safety Institute創設
- HUDERIA（Human Rights, Democracy and Rule of Low Risk and Assessment 人権、民主主義及び法の支配へのリスク・影響評価）登場

【新興国】

米英加・欧州委員会・オーストラリア・シンガポールなどは勿論だが、新興国（ウガンダなど）もAIマネジメントの国際標準化において頭角を現す。

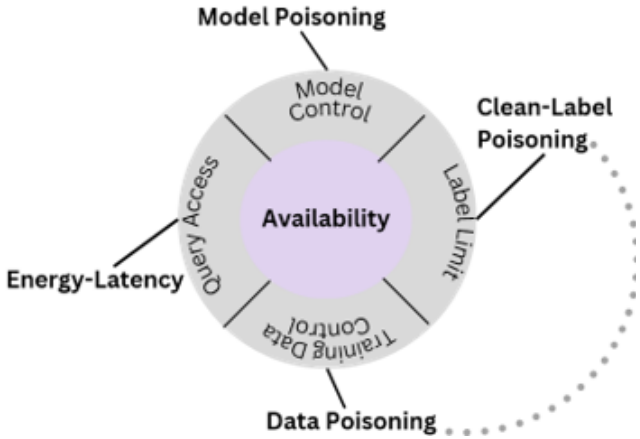
【日本】
- 産総研機械学習品質マネジメント
- AI Safety Institute創設
- AI事業者ガイドライン発行

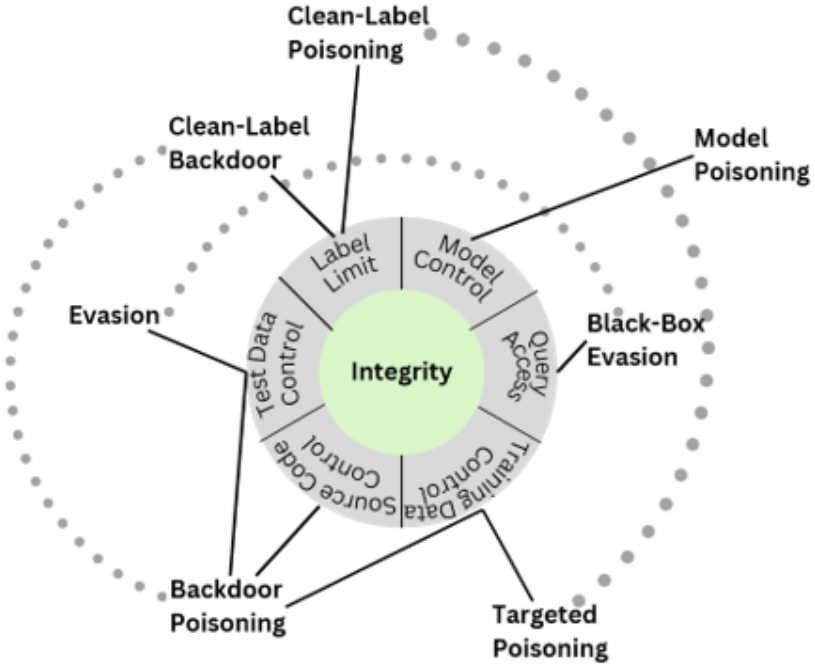
6.まとめ

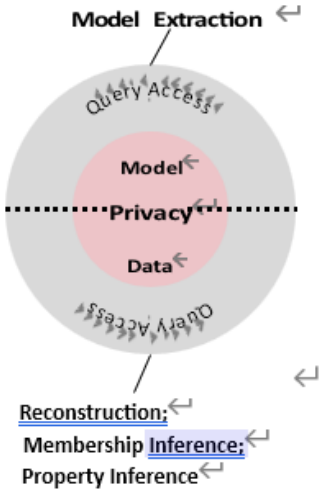
- 人工知能国際標準化シンポジウムを下敷きとし、米国を中心にAIマネジメントの世界的な動向を紹介した。
- AIに関する様々なリスクが産業や経済はもちろん、社会や安全保障をも揺るがす課題となっており、今後もAIマネジメントに向けた取り組みが求められる。

Appendix 米国情勢

PredAIの攻撃分類



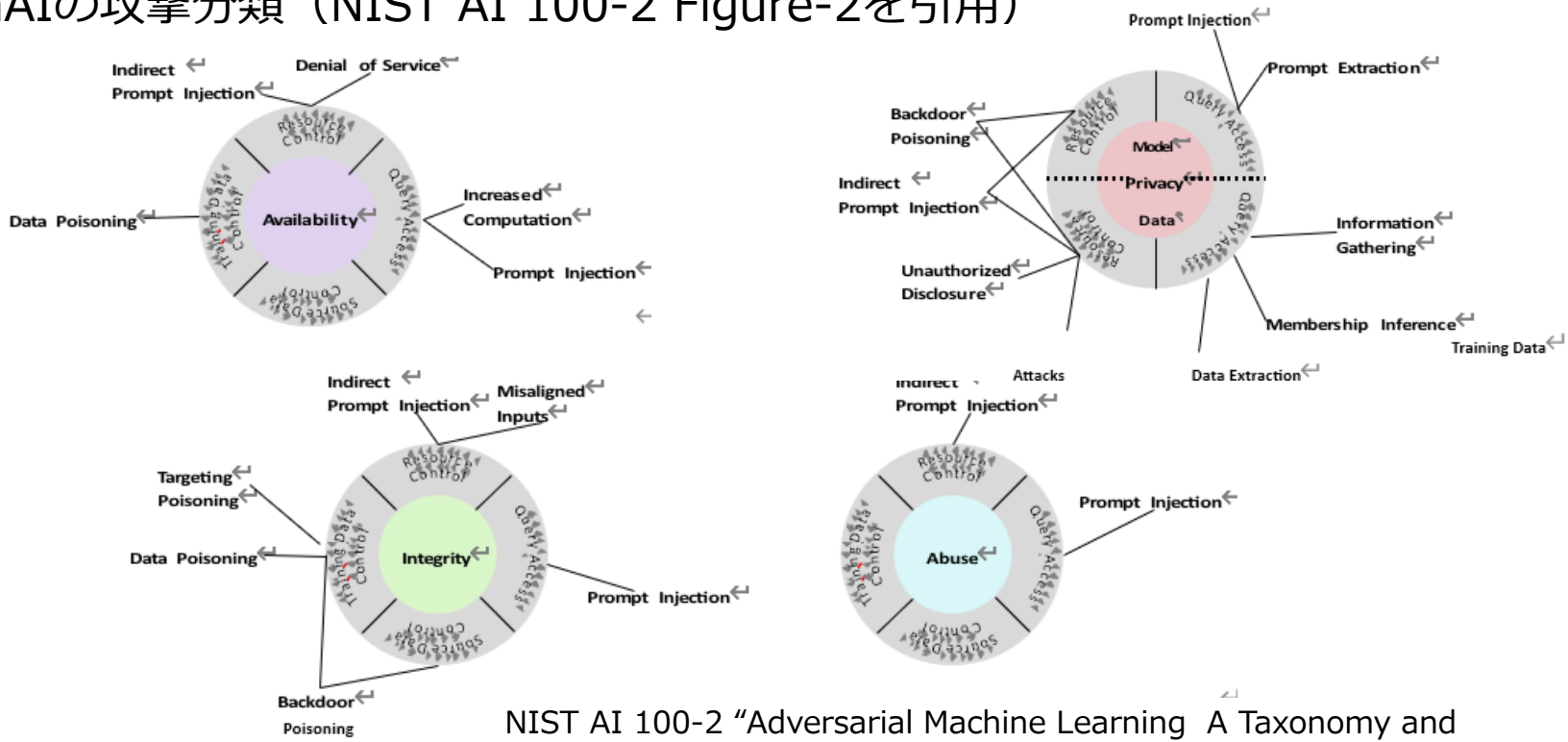




NIST AI 100-2 “Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations” Figure-1から引用

Appendix 米国情勢

GenAIの攻撃分類（NIST AI 100-2 Figure-2を引用）



NIST AI 100-2 “Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations” Figure-2から引用

Appendix 米国情勢

NIST AI RMFが規定するAI RMF Core概念

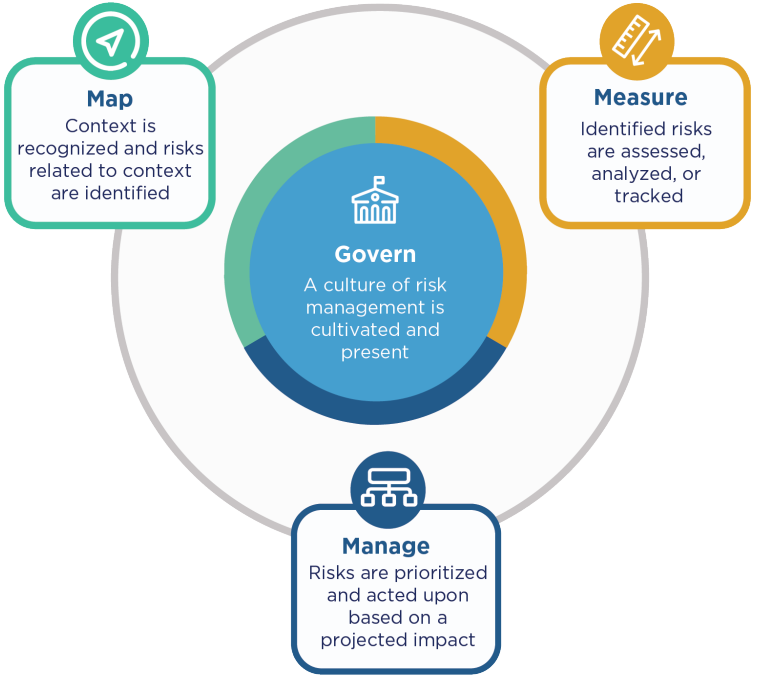
リスク管理のため求める機能(製品やサービスに求める機能ではなく、開発プロジェクトに求める機能)としてGOVERN, MAP, MEASURE, MANAGEを規定する。リスク対応のためのPDCAサイクル。

GOVERN：リスクマネジメントを実施する。組織の原則・方針・優先事項・原則まで決める。

MAP：開発対象に関係するAIアクター・潜在リスクの洗い出しや原因予測による改善プロセスの構築、開発対象のドメイン情報に至るまで徹底的に洗い出す。

MEASURE：AIリスクを分析、評価、ベンチマーク、監視する。そのためのツール・技術・手法を導入する。

MANAGE：MEASUREで測定されたリスクに対し、定期的に管理(GOVERN)しリスク対応のための資源を割り当てリスクに向けた対応を図る。



NIST AI 100-1 “Artificial Intelligence Risk Management Framework (AI RMF 1.0)” Figure-5から引用

Appendix 米国情勢

NIST AI 600-1が規定するリスク

CBRN Information(CBRN情報)

化学的(Chemical)、生物学的(Biological)、放射性物質に関する重大な悪意ある情報(Radiological)への参入障壁が低くなる、またはアクセスが容易になること。

Confabulation(詐称)：

自信満々に述べているが誤りまたは虚偽の内容（俗に「幻覚」または「捏造」と呼ばれる）を作り出すこと。

Dangerous or Violent Recommendations(危険または暴力的な推奨)

暴力的、扇動的、過激化的、脅迫的なコンテンツの作成とアクセスを容易にすること。

Data Privacy(データのプライバシー)

生体情報、健康情報、位置情報、個人を特定できる情報、またはその他の情報の漏洩、不正開示、匿名化解除、 位置情報、個人を特定できるデータ、またはその他の機密データの漏洩や不正開示、匿名化解除。

Environmental(環境)

GAIモデルのトレーニングにおけるリソースの大量使用による影響、および生態系に損害を与える可能性のある関連する結果

Human-AI Configuration(人間とAIの構成)

人間とAIシステムの配置や相互作用。アルゴリズム嫌悪、自動化への偏りや過度の依存、目標や望ましい結果の不一致や誤指定に基づくAIシステムによる欺瞞的または難解な行動。

Appendix 米国情勢

NIST AI 600-1が規定するリスク(続き)

Information Integrity(情報の完全性)

検証されていない可能性のあるコンテンツの生成と交換、および消費をサポートするための敷居が低くなり、事実と意見を区別せず、不確実性を認識しないこと。（CIAのIntegrityとは意味が異なる点に注意）

Information Security(情報セキュリティ)

セキュリティ攻撃、ハッキング、マルウェア、不正プログラム、不正アクセスの容易さなど、攻撃的サイバー能力の障壁が低下すること。

Intellectual Property(知的財産)

著作権、商標権、またはライセンス供与されたコンテンツが、無許可で使用されること。

Obscene, Degrading, and/or Abusive Content(わいせつ、下劣、虐待的なコンテンツ)

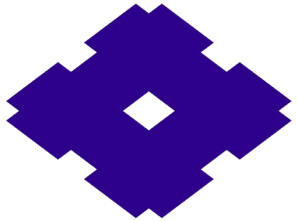
わいせつな、品位を欠く、または虐待的な画像の制作やアクセスを容易にした、 わいせつ、品位を傷つける、および/または虐待的な画像（合成児童性的虐待画像（CSAM）を含む）の制作とアクセスを容易にすること。成人の非同意的親密画像（NCII）

Toxicity, Bias, and Homogenization(毒性、偏見、均質化)

有害な、あるいはヘイトスピーチ、中傷的な、あるいはステレオタイプなコンテンツに公衆が曝されることを制御することが困難であること。

Value Chain and Component Integration(バリューチェーンとコンポーネントの統合)

GenAIによる自動化の進展により不適切に取得された、あるいはクリーニングされていないデータ、AIのライフサイクル全体にわたる不適切なサプライヤー審査、あるいは川下ユーザーに対する透明性や説明責任を低下させるその他の問題を含む。



住友電工

Connect with Innovation

https://sumitomoelectric.com/jp/

SUMITOMO
ELECTRIC
GROUP