AI 02

Beyond 5G | AI | 量子情報通信 | サイバーセキュリティ

外部連携可

基礎研究 応用研究 技術実証 社会実証

生成AIの安全性と信頼性

～ GPAI東京センターのSAFEプロジェクト ～

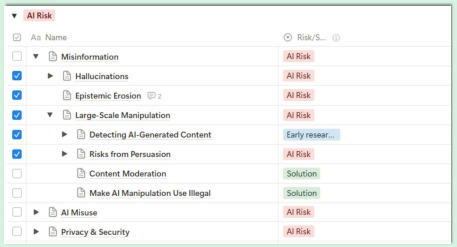
概要

SAFEプロジェクトは、世界初となるAI安全性リスクと対策の包括的マッピングを実施。政策立案者や開発者に向けた、実用的かつ直感的なナビゲーションツールを構築。東京センターは本プロジェクトの中核拠点です。

すべてのAI安全性リスクと解決策は、6つのトップレベルのカテゴリに分類

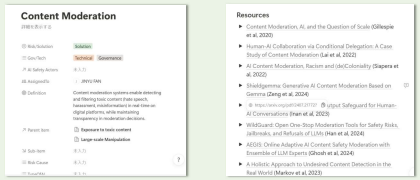
	カテゴリ名	例
1	悪用・犯罪リスク	サイバー攻撃など
2	差別・有害情報リスク	ヘイトスピーチ等
3	プライバシー・セキュリティリスク	個人情報漏洩など
4	システム障害リスク	重要インフラ停止など
5	偽・誤情報によるリスク	世論操作など
6	超知能による乗っ取りリスク	人間による制御の喪失など

各カテゴリについてリスクと解決策を確認



各項目について
参考文献を含む詳細な解説を表示

特定のキーワードによる検索や、特定の主体等が関係する項目だけをソートして表示させること等も可能



特徴

- グローバルなAI安全性リスクと解決策の体系的整理
- OECDやG7政策と連動した支援
- 直感的に探索できるデータベースを構築

ユースケース

- 規制当局による政策ギャップの特定
- 研究者による未解決課題の発見と共同研究先の探索
- 企業の開発現場でのAI安全性実装

今後の展開

- リスク・ソリューション記事数の大幅拡充
- 国際標準化と多地域間連携の深化
- プラットフォームへの専門家参加・貢献機能の実装

【お問合せ先】
GPAI東京専門家支援センター
Mail：gpai-tokyo-esc@ml.nict.go.jp

NICT オープンハウス 2025