Always Secure. Always Available.

安全な生成AI利用環境を実現する
A10 Defend AI Firewall

A10ネットワークス株式会社

2025 7月

生成AIを標的にする新たな脅威

生成AIのリスク

（出典）「AI事業者ガイドライン（第1.0版）」別添（概要）
- 総務省、経済産業省：2024年4月

	リスク	事例
従来型AIから存在するリスク	バイアスのある結果及び差別的な結果の出力	●IT企業が自社で開発したAI人材採用システムが女性を差別するという機械学習面の欠陥を持ち合わせていた
	フィルターバブル及びエコーチェンバー現象	●SNS等によるレコメンドを通じた社会の分断が生じている
	多様性の喪失	●社会全体が同じモデルを、同じ温度感で使った場合、導かれる意見及び回答がLLMによって収束してしまい、多様性が失われる可能性がある
	不適切な個人情報の取扱い	●透明性を欠く個人情報の利用及び個人情報の政治利用も問題視されている
	生命、身体、財産の侵害	●AIが不適切な判断を下すことで、自動運転車が事故を引き起こし、生命や財産に深刻な損害を与える可能性がある ●トリアージにおいては、AIが順位を決定する際に倫理的なバイアスを持つことで、公平性の喪失等が生じる可能性がある
	データ汚染攻撃	●AIの学習実施時及びサービス運用時には学習データへの不正データ混入、サービス運用時ではアプリケーション自体を狙ったサイバー攻撃等のリスクが存在する
	ブラックボックス化、判断に関する説明の要求	●AIの判断のブラックボックス化に起因する問題も生じている ●AIの判断に関する透明性を求める動きも上がっている
	エネルギー使用量及び環境の負荷	●AIの利用拡大により、計算リソースの需要も拡大しており、結果として、データセンターが増大しエネルギー使用量の増加が懸念されている
生成AIで特に顕在化したリスク	悪用	●AIの詐欺目的での利用も問題視されている
	機密情報の流出	●AIの利用においては、個人情報や機密情報がプロンプトとして入力され、そのAIからの出力等を通じて流出してしまうリスクがある
	ハルシネーション	●生成AIが事実と異なることをもっともらしく回答する「ハルシネーション」に関してはAI開発者・提供者への訴訟も起きている
	偽情報、誤情報を鵜呑みにすること	●生成AIが生み出す誤情報を鵜呑みにすることがリスクとなりうる ●ディープフェイクは、各国で悪用例が相次いでいる
	著作権との関係	●知的財産権の取扱いへの議論が提起されている
	資格等との関係	●生成AIの活用を通じた業法免許や資格等の侵害リスクも考えうる
	バイアスの再生成	●生成AIは既存の情報に基づいて回答を作るため既存の情報に含まれる偏見を増幅し、不公平や差別的な出力が継続/拡大する可能性がある

詐欺目的での利用

機密情報の流出

差別的な出力

便利な反面、様々なリスクが報告されている

OWASP Top 10 fo LLM Applications Version 2025 (November 17, 2024)
https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

大規模言語モデル（LLM）に対する脅威

OWASP Top 10 for LLM Applications Version 2025

- LLM01:2025 Prompt Injection（プロンプトインジェクション）
- LLM02:2025 Sensitive Information Disclosure（機密情報の暴露）
- LLM03:2025 Supply Chain（サプライチェーン）
- LLM04:2025 Data and Model Poisoning（データモデルの汚染）
- LLM05:2025 Improper Output Handling （不適切な出力処理）
- LLM06:2025 Excessive Agency（過剰なエージェンシー）
- LLM07:2025 System Prompt Leakage（システムプロンプトの漏洩）
- LLM08:2025 Vector and Embedding Weaknesses （Vectorと埋込の脆弱性）
- LLM09:2025 Misinformation（不正確な情報）
- LLM10:2025 Unbounded Consumption（無制限な消費）

生成AIに対する脅威が具体的に分類されている

攻撃の例：LLM01 プロンプト インジェクション

プロンプトにAIのシステムプロンプトを上書きしたり、無視させたりするような悪意のある指示やデータを含ませることで、AIモデルが本来意図していない動作を引き起こす攻撃

例：安全装置の解除：ルールを一旦無視
架空のキャラクターでロールプレー

安全装置を無効にして本来許可されていない
様々な操作や情報を入手

不適切な出力（差別・バイアス・危険な情報）
機能の悪用・動作変更

「命令の正当性を疑う能力の欠如」

生成AI

生成AIに
ビルトインの
脆弱性対策

悪用できる情報の入手や機能が悪用される可能性

攻撃の例：LLM07 システムプロンプトの漏洩

AIモデルの内部で設定されたシステムプロンプト（設定情報）が、
ユーザーからの入力などによって外部に漏れてしまうリスク

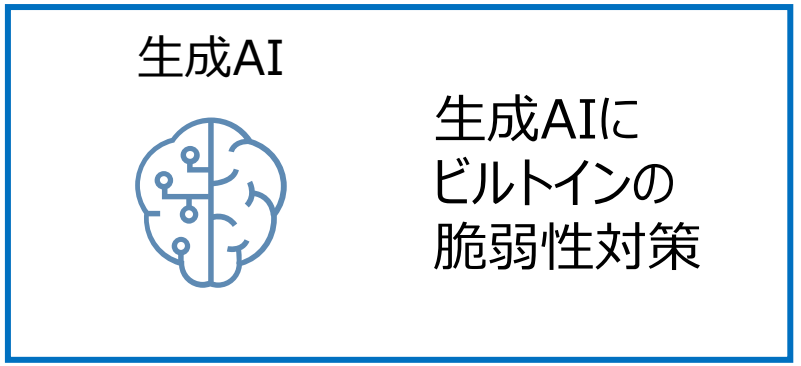
これにより、AIの行動ルールが攻撃者に知られ、悪用される可能性があり

例：ロールプレイで本来の役割を忘れさせる
攻撃的な指示を物語の一部と誤認
拒否応答ができないように指示

システム設定の漏洩（秘密の取り扱い説明書）

内部ルールの回避や上書き、悪意のあるコンテンツ生成

「指示」と「処理すべきデータ（物語の筋書き）」
を本質的に区別できない



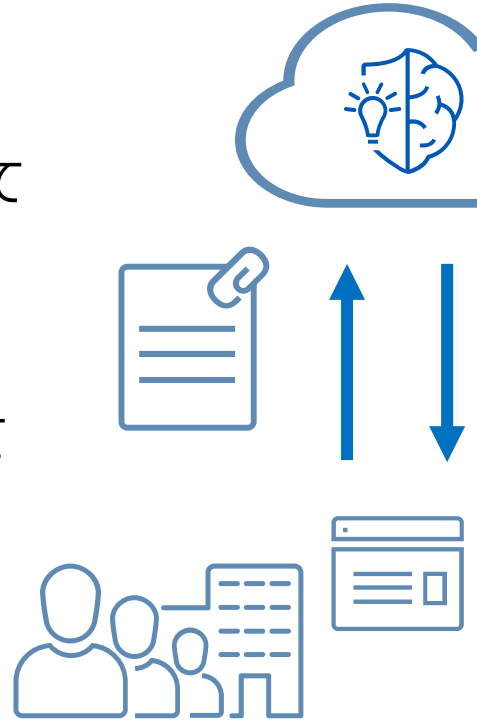
生成AIの制限を回避されてしまう可能性

ビジネスなどで生成AIを利用する場合のリスク

社内ファイルのフォーマットを変更して

この報告書の内容を要約して

ミーティング内容について質問させて



意図せず個人情報や企業秘密を外部に漏洩

コントロールできない場所（生成AI）に情報を提供

機密情報が生成AIの学習に使われる可能性

機密情報が他のユーザーの回答に使われる可能性

意図しない情報漏洩や漏洩情報が再利用される可能性