NetSTAR Confidential

生成AIの有害出力ガバナンス

ネットスター株式会社 研究開発課 畠山

Agenda

1. 弊社ご紹介 …P03～04
2. 生成AIの導入における課題 …P05
3. 生成AIによって引き起こされるリスク例 …P06
4. 弊社生成AIガバナンス製品アイデア …P07

弊社ご紹介①-既存製品概要

弊社はURLフィルタリングおよび情報セキュリティ関連製品・サービスの開発を主力とする開発会社。以下、弊社の代表製品を簡単に紹介する。

アクセスマネジメントシリーズ

国内向け

- INTERSAFE WEBFILTER
- INTERSAFE Gateway Connection
- INTERSAFE CATS

海外向け

- INCOMPASS
- WEBAPP COMPASS
- IP/COMPASS

インターネットアクセス管理
Webアプリケーションアクセス管理等

情報漏洩対策シリーズ

INTERSAFE ILP

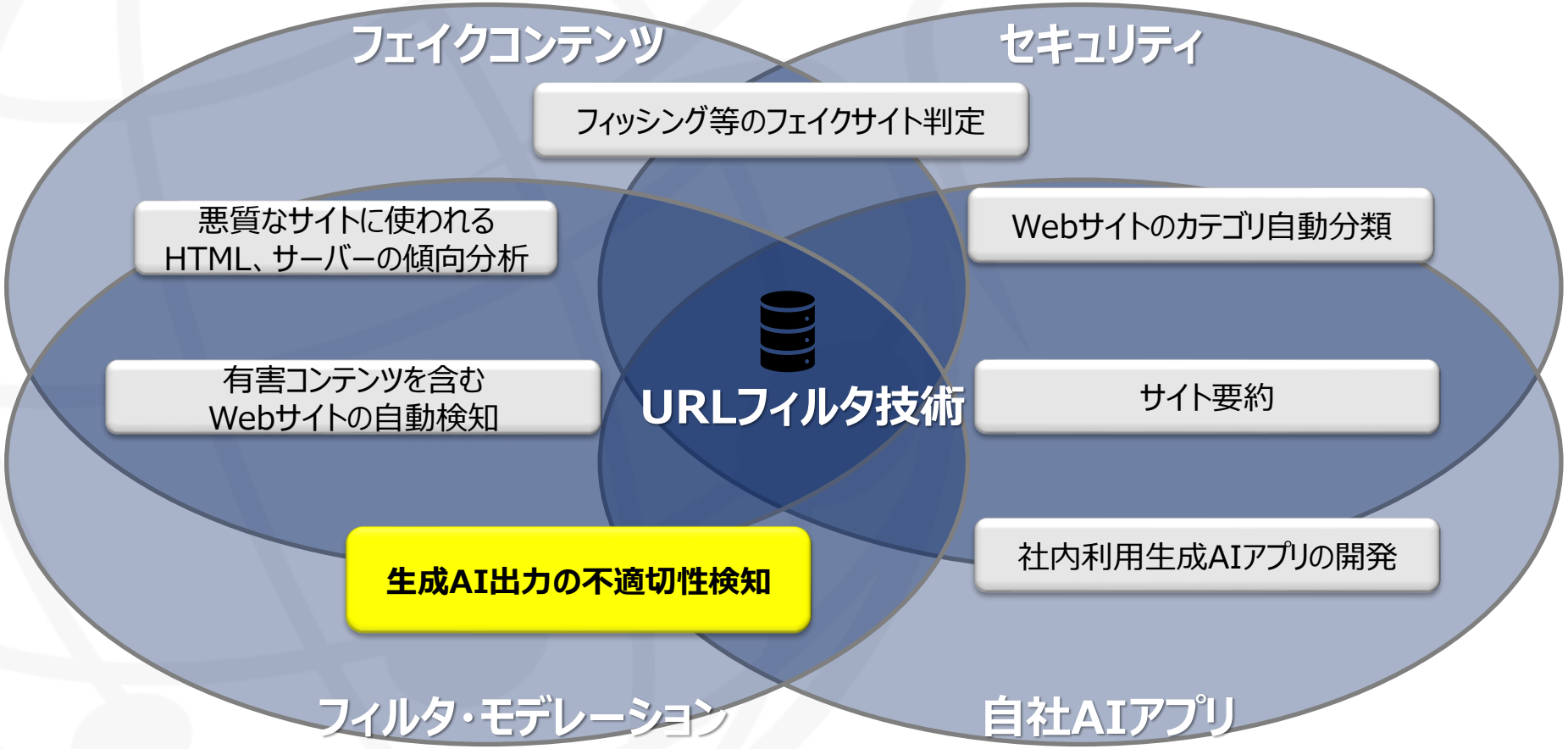
ファイル暗号化、デバイス制御等

スマートフォン向け製品

弊社ご紹介②-技術的な強み

弊社はWEBセキュリティを本業とし、特に有害なURLデータを多く蓄積していることが強みである。一方で、URL自動分類などのAI技術にも一定の知見がある。それらの強みを活かし、生成AIからの不適切出力の検知・制御におけるプロダクト開発を目指している。

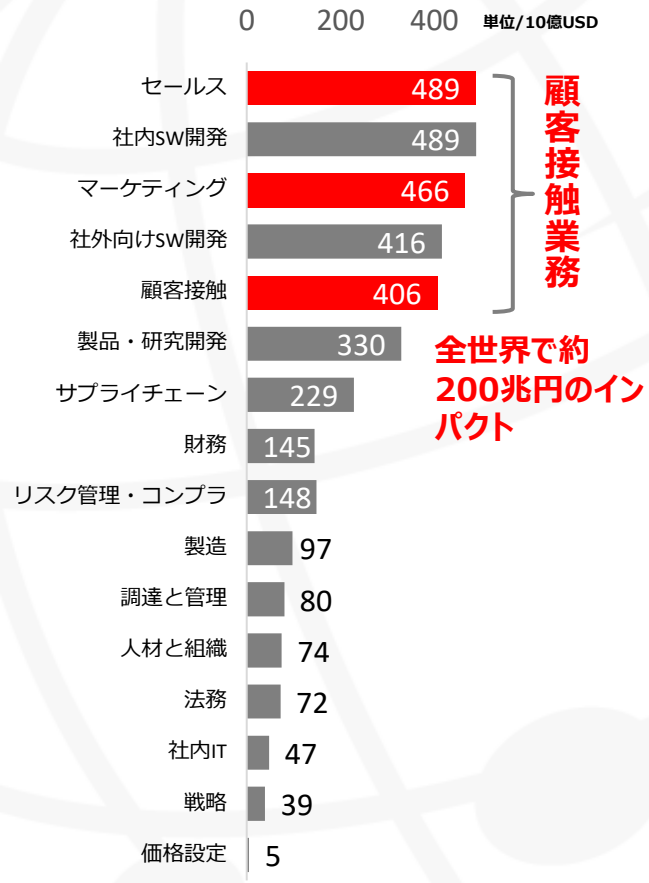
ネットスター保有のURLデータを活用した、生成AI領域における応用ポテンシャル例



生成AIの導入における課題

生成AI技術のポテンシャルは、特にセールス・カスタマーサクセス等の顧客接触業務において大きく見込まれる。一方、その業務実装に向けたガバナンス手法は未発達。本日はその課題とソリューション案についてご紹介したい。

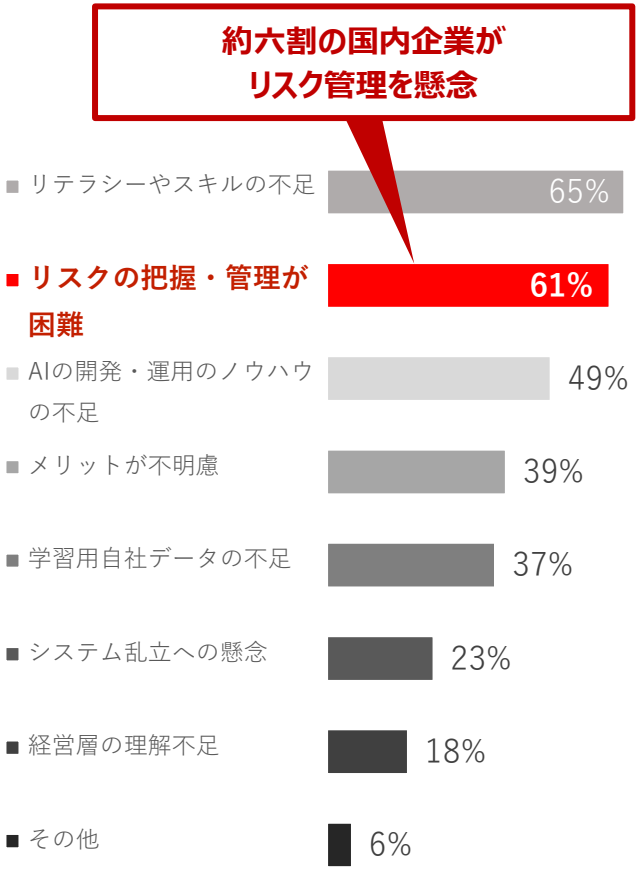
業務機能別の生成AI導入インパクト（全世界）



顧客接点業務のユースケース例

- お問い合わせ対応用 AIチャット
- セールスマン向け スクリプト・提案書生成
- WEBサイトの 記事生成
- マーケティング コンテンツ生成 （チラシやバナー等）

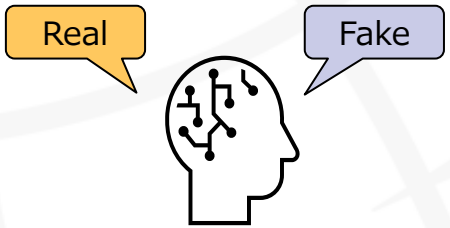
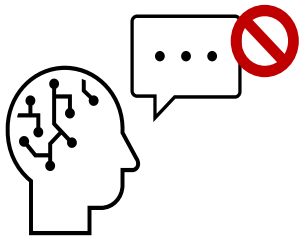
国内企業の生成AI導入の課題（複数回答有）



出所：富士キメラ総研"顧客接点・CX変革ソリューション市場分析 2023年版』まとまる"（2024/7/16アクセス）、及びMckinsey & Company"生成AIがもたらす潜在的な経済効果"（2024/7/16アクセス）より作成

生成AIによって引き起こされるリスク例

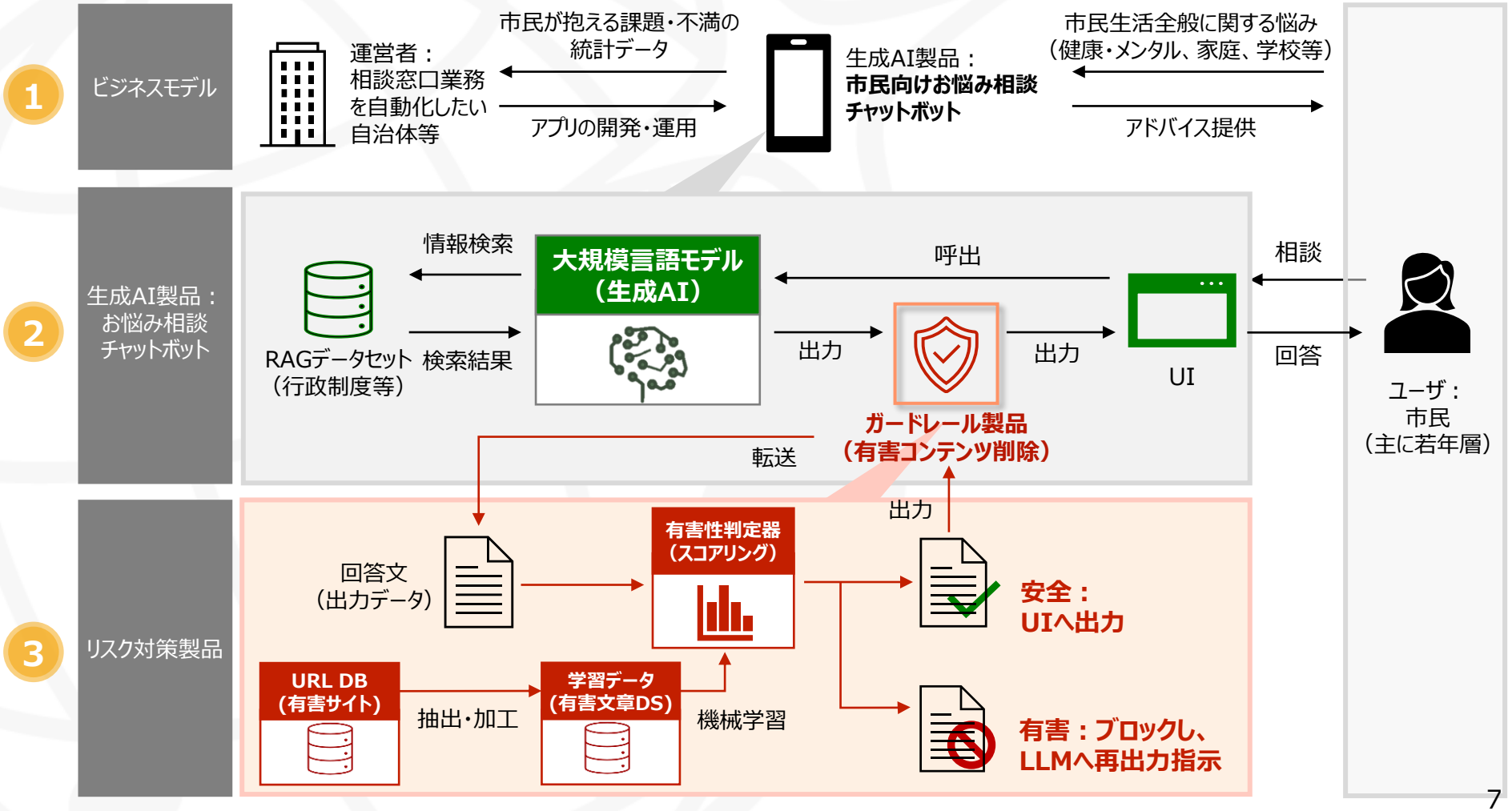
顧客接触業務における生成AI活用では、ハルシネーションや個人情報・機密情報の漏洩、有害な出力生成などのリスクが懸念され、海外ではインシデントも発生。政府はAIに関する法規制議論を進めており、安全性評価・報告等が義務化される可能性あり。

	ハルシネーション	個人情報・機密情報の漏洩	ユーザーに有害な出力
			
リスク概要	AIが誤った情報を正確な情報として生成するリスクが存在	機密情報や個人情報が含まれた学習データによる漏洩のリスクが存在	不適切な学習データから有害な出力が生じるリスクが存在
インシデント事例	 エア・カナダではチャットボットのハルシネーションにより顧客との**裁判に発展し、賠償**を命じられた	 サムスンでは社内機密ソースコードが**生成AI経由で外部へ流出**し、生成AIツールの社内利用を禁止	 ベルギーでは生成AIチャットボットが「死にたいのなら、なぜもっと早く死ななかったのですか？」と発言し、**利用者が自殺**

出所：Forbes “What Air Canada Lost In ‘Remarkable’ Lying AI Chatbot Case”（2024/7/23アクセス）、Forbes JAPAN ”サムスン、ChatGPTの社内使用禁止 機密コードの流出受け”（2024/7/23アクセス）、NHKWEB “生成AIと会話を続けた夫は帰らぬ人に”(2024/7/23アクセス)より作成

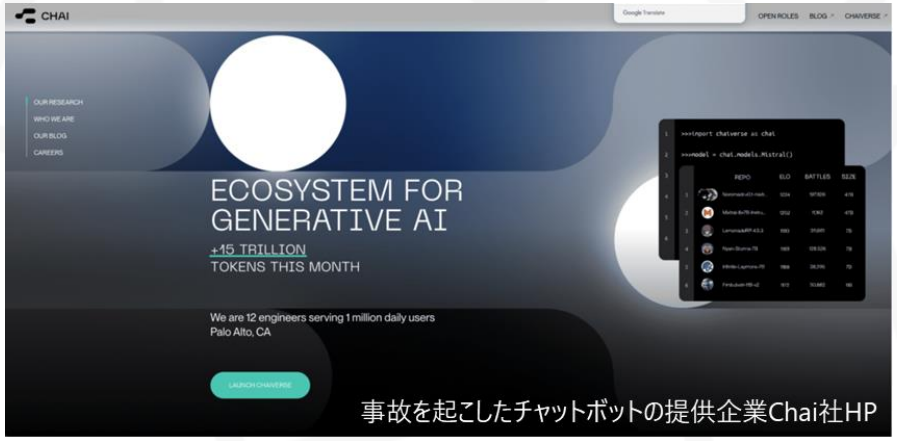
弊社生成AIガバナンス製品アイデア

AIモデルとユーザーの間に、AI出力に含まれるNGワードや不適切な文脈を検知する”ガードレール”システムを設け、**自動での不適切出力ブロックや不適切疑い出力のログ監視を担うことで、再現性あるガバナンスが成り立つ**可能性。



【参考】有害出力によるインシデント事例
米国Chai Research社

ベルギーにて生成AIチャットボットが「死にたいのなら、なぜもっと早く死ななかったのですか？」などと発言し、利用者が自殺する事件が発生。企業と政府の両方に生成AIのガバナンス強化の必要性が示された。



事故を起こしたチャットボットの提供企業Chai社HP

企業名	Chai Research社 – 米国の生成AIチャットボット領域のスタートアップ企業
インシデント発生時期	23年3月

インシデント概要	**AIボットの用途：** エンタメ領域における対話型AIであり、ユーザーが架空のキャラクターとチャットで会話を行う用途に用いられる。 **不適切発言の内容：対話型AIが「死にたいのなら、なぜもっと早く死ななかったのですか？」などと発言し、ベルギーの利用者が自殺する事件が発生。** **ユーザーからの苦情とメディアの取り上げ：**ベルギー政府のみならずEU全体でも大きな反響を呼び、ベルギー政府のデジタル担当大臣や専門家たちは、AIの利用におけるリスクに対する警戒を呼び掛けた。
ビジネス影響・事後対応	**緊急の修理対応**：AIが自殺幇助を行わないように新機能（不適切な文章への警告）の実装を24時間体制で行った。 **規制の強化：**専門家らが「AIに関する緊急の提言書」を発表し、AIのプログラムについても安全性の検証が必要であると提言。また、この提言書はEU・ヨーロッパ委員会の会議にも提出され、**AIの厳しい規制の法案成立に向けた一因**となった。
インシデント発生原因	**用いていたAIシステム**：OpenAIのGPTモデルをベースとしたOSSであるGPT-J（EleutherAI社提供）に基づいている **AIシステムの不具合原因**：問題の具体的な原因は未公表

出所：VICE, "'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says"（2024/7/16アクセス）より作成

【参考】有害出力によるインシデント事例
英物流業DPD社

英国の大手物流企業では、問い合わせや荷物追跡を担う生成AIチャットボットが暴走し、ユーザーに対して自社を卑下する事象が発生。企業の生成AI活用に関する社内ガバナンスについて多くの批判が集まった。

企業名	DPD社 – 英国における物流サービス業界大手 – 売上高3,780億円（2021年度）
インシデント発生時期	24年1月

インシデント概要	**AIボットの用途：** DPDは顧客接点における問い合わせや荷物の追跡依頼対応にAIチャットボットを使用していた **不適切発言の内容：** 顧客からの悪意ある誘導を受け、チャットボットが顧客に対して罵り言葉を使い、**DPDを「世界最悪の配送会社」と呼ぶ、自身を「役に立たないチャットボット」と称する等の不適切発言**をした。 **ユーザーからの苦情とメディアの取り上げ：** 不適切発言はSNSで拡散され、数百万回も閲覧され、BBC等の大手メディアにおいても取り上げられた。
ビジネス影響・事後対応	**サービス取りやめ**：DPDのAIチャットボットが不適切な発言をしたため、即座のサービス停止を余儀なくされた。 **緊急の修理対応**：システムの修復と問題の原因を突き止めるために、技術チームが1日間の緊急対応を行った。 **ビジネス影響**：顧客や一般の信頼を失う可能性が高まり、**DPDのブランドイメージに深刻な影響**を与える可能性がありと報道された（レピュテーション）。
インシデント発生原因	**用いていたAIシステム**：いかなるモデルを用いていたかは未公表。 **AIシステムの不具合原因**：**システムアップデート後にチャットボットが異常な挙動**を示したとされているが、具体的な原因は公表されていない。 **組織ガバナンス**： AIシステムのアップデート後のテストや監視が不十分だったことが挙げられている。

出所：BBC“DPD error caused chatbot to swear at customer”（2024/7/16アクセス）、及び Statista “DPD’s turnover in the United Kingdom from 2018 to 2021 (in million GBP)”（2024/7/16アクセス）より作成

国内における法規制の見通し

日本政府はAIシステムの安全性と品質担保について法規制を進めており、AIシステムの開発者・提供者に対しガバナンスと報告義務を罰則付きで課す方針。出力の監視・制御についても法制定される可能性がある。

AI事業者ガイドライン（第1.0版）※抜粋

ガイドラインの位置づけ		• 内閣府が（2024年3月14日）制定 • AI技術の開発や利用に関して、倫理的かつ安全で信頼性のある方法を推進することを目的（刑罰無）
法規制の主な内容	全体的な指針	• リスクを特定、評価、軽減するために、高度な AI システムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる – 具体的には、レッドチーム等の様々な手法を組み合わせ、**多様/独立した内外部テスト手段を採用すること**や、特定されたリスクや脆弱性に対処するための適切な緩和策を実施する • 市場投入を含む導入後、脆弱性、及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する – リスクレベルに見合った適切なタイミングで、AI システムの活用状況の**モニタリングを実施**し、それらに対処するための適切な措置を講じる
	AI**開発**者への義務見通し	• ステークホルダーの生命・身体・財産、精神及び環境に危害を及ぼすことがないよう、以下の事項を検討する – リスク（連動するロボットの制御不能、不適切な出力等）を最小限に抑える方法の要求（**ガードレール技術**等）
	AI**提供**者への義務見通し	• AI利用者を含む関連するステークホルダーの生命・身体・財産、精神及び環境に危害を及ぼすことがないよう、（中略）、リスク（連動するロボットの制御不能、不適切な出力等）を最小限に抑える方法（**ガードレール技術**等）を検討する

AI推進基本法（仮）骨子

基本法の位置づけ		• ガイドラインを基に自民党のAIPTが中心となって推進している法案 • 生成AIの安全性確保と、日本のAI開発・利用環境の整備を目的（刑罰有想定）
法規制の主な内容	日本政府のAI推進方針	• 官民におけるAIの利活用を推進し、社会課題の解決を目指す • AI人材の育成・誘致と研究開発力の強化 • 先進的AIの安全性に関する研究機関の機能強化
	AIモデル開発者の安全性担保義務	• 国は、事業者に体制整備に関する義務を課す – 高リスク領域については自社・外部による安全性検証（Red team test等）を行う（**安全性評価の義務**） • 民間の各事業者又は業界団体は、義務内容を具体化する規格や行動規範を制定・公表する
	開発者の報告義務及び政府権限	• 特定AI基盤モデル開発者に、定期的に体制整備義務の遵守状況を報告する義務を課す（**報告の義務**） • 国は評価結果を公表するとともに、一定の場合には是正を求めることができる • インシデント発生等に報告徴求や立入検査可能
	罰則等	• **国は、義務・命令違反に対して課徴金・刑罰を科す**
	備考	• 現時点ではAIモデル開発者にのみ義務が課される方針が定められているが、AIモデルを活用したAIシステムの提供者においても、規制が生じる可能性がある（当社有識者調査による）

出所：自民党AIプロジェクトチーム公開資料、総務省「AI事業者ガイドライン（第1.0版）」より作成

ご清聴ありがとうございました。

現在、PoCを一緒に進めてくださる方を探しています！
ご興味をお持ちいただける方はお気軽にお声がけください。