TOKIO MARINE

リスクマネジメント最前線

2023 | No.5

AI を利活用するために考えるべきリスクマネジメントとは

ビジネスリスク本部　研究員　加藤　直人
専門分野：製造業における営業データ分析、海上輸送コンテナ積載シミュレーションシステムの設計開発等に従事。統計学、データ分析を活用したリスクマネジメント体制構築、製造業に対する事業継続計画策定等を支援。

ビジネスリスク本部　主任研究員　木村　圭佑
専門分野：製造業における製品開発・品質保証に従事。製造業、建設業、小売業、金融・保険業、広告業に対する事業継続計画策定、災害対応訓練、リスクマネジメント体制構築等を支援。

昨今、ChatGPT[1]のような生成型 AI（Generative AI）[2]が注目されたことを契機に、企業・自治体等にて AI 技術を業務に活用する動きが一層進んでいる。同時に、個人情報の漏洩や安全性の問題等、AI 活用のリスクも改めて浮き彫りになっており、例えば G7 群馬高崎デジタル・技術大臣会合では AI ガバナンスのグローバルな相互運用性を促進等するためのアクションプランに合意がなされた[3]ように、様々な国や組織で AI システムに係るリスクマネジメントの手法・フレームワークを定めようとする動きがある。

本稿では、米国の NIST（National Institute of Standards and Technology）という機関により公表された「AI Risk Management Framework 1.0」[4]を参考にしつつ、AI システムを利活用するためのリスクマネジメントについて提言する。

1. 近年の AI を取り巻く状況について

（１）近年の AI 発展と企業・自治体等の動向

AI は、加速度的に発展しており、深層学習（ディープラーニング）といった機械学習技術に代表されるように、自動運転、需要予測、外観異常検査等、様々な業界で AI 技術の応用が進んでいる。特に、近年では画像生成 AI「Midjourney[5]」「DALL・E 2[6]」や文書生成 AI「ChatGPT」「Bard[7]」「Perplexity AI[8]」等の生成型 AI がリリースされて以降、文章で AI システムに指示を出せる手軽さもあり、技術者に限らず一般に利用者が広まりつつある。そうした社会の動きに応じて、Google や Amazon などの大手テック企業は生成型 AI 市場への参入を発表し[9] [10]、その他の企業・自治体等においても業務改善や広報活動のために ChatGPT の機能を導入する動きがみられるようになった（図表１）。

■　図表１　企業・自治体等で AI を導入・検討した事例

組織名	概要、見込まれる効果
パナソニック ホールディングス株式会社	DX プロジェクト「Panasonic Transformation（PX）」の一環としてパナソニックグループの国内全社員 9 万人向けに Azure OpenAI Service を組み込んだ AI アシスタント「PX-GPT[11]」を提供。様々な部門のビジネスアイデアの創出を図っている。

株式会社ベネッセホールディングス	1 万 5 千人の社員を対象に、Azure OpenAI Service を組み込んだ独自の AI チャット「Benesse GPT[12]」を開発したことを発表。将来的には消費者向けでのサービス展開を視野に入れている。
三井化学株式会社	日本 IBM との協業により、既存のデータマイニングアプリに Azure OpenAI Service を組み込むことで、新規用途探索のための利便性向上を図っている[13]。
茨城県	県の公認バーチャル Youtuber「茨ひより」に ChatGPT の機能を組み込んだ「AI 茨ひより」を、2023 年 4 月末開催の千葉市内のイベントにて公開。茨城県の PR 等を行った[14]。
神奈川県横須賀市	自治体専用ビジネスチャットツールに ChatGPT の API[15]を連携。文章要約・文章作成・誤字脱字チェック、アイデア創出等に活用できるようにすると言及[16]。
農林水産省	すでに公表された文書の更新作業について、ChatGPT を用いてわかりやすく表現・簡素化することを検討[17]。

出典：各種報道機関・公開情報をもとに弊社作成

（2）AI の有用性と限界

様々な AI 技術が過去から活用されてきている。例えば、予測や分類を行う AI は故障診断や需要予測等に応用されており、画像認識・音声認識を行う AI は自動運転や異常検査等に利用されている。従来の AI システムの多くは特定の値や画像等を入力データとし、その分類や予測値を出力データとして返すという識別機能を果たしてきた。しかしながら、それらの結果を利活用するには専門的な知識が要求されるため、利用者は主に専門家であり、機能も特化したものが大半であった。比較的一般に用いられるものとしては、スマートフォンに搭載されている生体認証や自動運転があげられる。

一方、生成型 AI では識別から一歩進み、入力された文章に応じて創造的な画像や文章を出力することが可能となった。操作についても比較的簡素であることから、今後も多くの一般人が AI 技術に直接触れることが考えられる。生成型 AI には様々な利用例が挙げられるが、一般的に文章作成が業務負荷になることが多いため、文書作成 AI である ChatGPT を利用する動きが多くみられる。例えば、図表 1 に示したとおり、農林水産省は公表済みの各種文書の要約・簡素化の補助手段として用いることを検討している。

なお、留意すべき事項として、ChatGPT が誤った情報を出力するケースも散見されている。例として、ChatGPT に「東京特許許可局について教えてください」という質問をしたところ、「特許庁の管轄の下、東京都千代田区霞ヶ関に所在している」という旨の誤った返答をすること（2023 年 5 月上旬時点）等が挙げられる[18]。これは ChatGPT の背景にある大規模言語モデルが、打ち込まれた文章を読み取り、連想される内容を文法上問題ない文章にまとめた結果であり、東京特許許可局は早口言葉として有名ではあるが実在しない機関である。なお、同言葉は ChatGPT が興味深い返答をする事例として注目を集めている。このような ChatGPT の出力の仕方をイメージするうえでは、スマートフォンに搭載されている予測変換機能を参考とされたい。例えばスマートフォンに単語を打ち込むとそれに続くであろう単語が表示され、日常的な定型文であれば予測変換機能のみで文章を書き切れることもあるが、大抵はだんだんと意味の通じない文章となるだろう。一方、ChatGPT で用いられている大規模言語モデルは文章全体に注意を向けながら次の言葉を予測してつなげることができるため、話に脈略をつけながら尤もらしい語順で言葉を紡いでいくことができる。しかし、文法上問題がなくとも先述の通り誤った内容を出力することがあるため、利用には注意が必要といえる。

以上のように、AI には有用性とリスクがある。AI 活用がこれまで以上に一般化することが予想される中で、企業や自治体としても的確にリスクを捉えながら有効活用を進めることが求められる。

（3）AI 活用によるインシデントとその分類

先述のとおり、生成型 AI によるアイデアの創出や自動運転による利便性の向上等様々な効果が期待される一方で、AI 活用によるリスクも着目されるようになった。実際、自動運転車の事故やディープフェイク[19]など、生成型 AI に限らず AI 全般に関連するインシデントがこれまでにも多く発生している。

様々な様態の AI リスクを管理していくには、先ず AI によってどのようなインシデントが発生しているのかを理解し、それらが及ぼした影響に応じて分類していくことが一つのヒントになると考える。本項では、米国の代表的な IT 関連企業等によって組織されている非営利団体「Partnership on AI[20]」によって集約・公表されているインシデント[21]を一部参考に、それぞれについて考えられる影響分類とその事例を以下のようにカテゴライズした（図表２）。

■ 図表２ AI 関連インシデントの分類

影響分類		インシデント例
AI システム由来	安全性	チェスの大会で、対局用ロボットが対戦相手の少年の手をつかみ、指を骨折させる事故が発生。
		自動運転をしている車がトレーラー車と衝突する事故が発生。自動運転側のドライバーが死亡。
	セキュリティ	ChatGPT を操作することでマルウェアを作成できることが研究者により発見。
		銀行の音声認識で電話口の顧客を識別するシステムにおいて、当人の双子の声で誤認証が発生。
	説明可能性	従業員が、AI システムの判断によって急遽解雇された事例が発生。
		本人の合意なしに AI を活用して算出した内定辞退率データを販売。購入した企業は選考の合否に用いていないと弁明したがその実情は不明。
	プライバシー	SNS 上の顔認証機能に、ユーザーから無断で得た顔写真をデータセットに用いているとして、プライバシーの侵害として訴訟が発生。
		商品の注文を受ける AI チャットボットがユーザーの同意無しに音声データを収集していたとして、プライバシーの侵害として訴訟が発生。
	公平性の欠如（バイアス）	リリースした写真編集アプリで、黒人の顔写真が「ゴリラ」と認識される不具合が発生。
		社員採用に用いていた AI に、女性に関連するワードがあると評価が下がる仕様が判明し、運用を中止。
人間のふるまい由来	悪用	AI により生成された、ある企業の CEO に似せた音声を用い、関係者から大金をだまし取る事件が発生。
		ある国の要人が逮捕されたというディープフェイク画像が SNS 上で拡散される事案が発生。
	ヒューマンエラー	実証実験にて、自動運転バスがガードレールに接触。原因は車両内の自動運転に係るパーツの再起動忘れに原因があったと公表。
		社員が ChatGPT に機密情報・ソースコードを入力する事案が発生。

出典：「Partnership on AI」に公表されている情報・各種報道機関をもとに弊社作成

影響分類については以下の通り整理を行った。

□AI システムに由来する影響分類

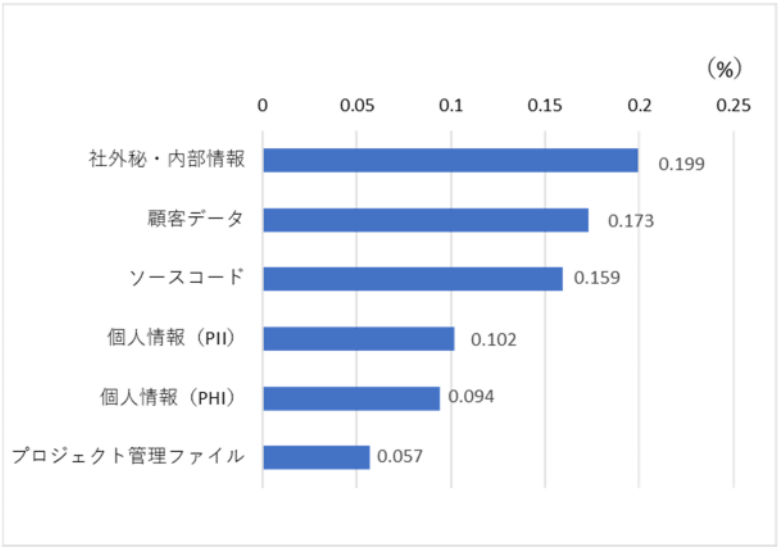
- 安全性：特に使用者の生命・健康等が危ぶまれるインシデント
- セキュリティ：主にサイバー攻撃に関連するインシデント
- 説明可能性：AI の判断プロセスや根拠が不透明であることに起因したインシデント
- プライバシー：AI システムのデータセット・入力に個人情報・機密情報を用いたインシデント
- 公平性の欠如（バイアス）：AI システムの不適切な出力によるインシデント

□組織や人間のふるまいに由来する影響分類

- 悪用：人間の悪意などに基づく意図的な対応に関連するインシデント
- ヒューマンエラー：人間の不注意等のヒューマンエラーに関連するインシデント

なお、「人間のふるまい」の観点で興味深いデータがある。Cyberhaven Labs[22] から公表されている Cyberhaven 製品を使用する企業の従業員 160 万人の ChatGPT の使用状況に関するデータ[23]によると、従業員の 8.2%が「少なくとも一度は職場で ChatGPT を使用した」とあり、業務への生成型 AI の活用が進んでいることが分かる。一方で、従業員の 3.1%が「会社の機密情報を ChatGPT に入力した」というデータも得られた。また、同調査で 2023 年 2 月 26 日～3 月 4 日の 1 週間に取得したデータ（図表 3）からは、従業員の約 0.20%が社外秘・内部情報を、約 0.17%が顧客データを、約 0.16%がソースコードを入力していたことも判明した。

■ 図表 3 2023 年 2 月 26 日～3 月 4 日の期間における機密情報入力の割合



出典：Cyberhaven Labsが公表するデータより弊社作成

このようなデータを勘案すると、AI システムに由来するリスクについてはもちろんのこと、従業員の悪用あるいはヒューマンエラー（無知に紐づくものも含む）により、企業・自治体等に損害を与える可能性についても看過できないリスクであると考える。

AI 導入により新たなアイデア創出、工数削減等のメリットが見込まれるため、今後も組織で AI を活用する、または導入を検討・要望する動きはさらに活発になることが想定される。しかし同時に、AI を用いることで情報漏洩や安全等に関するリスクにさらされる可能性もありうるということである。企業・自治体等としては AI の活用や導入における懸

念を払拭するためにも、AI に係るリスクを知り、管理態勢を構築していくことが益々重要になっていくと考えられる。次の第 2 章では、AI のリスクマネジメントに関するフレームワークについて紹介し、第 3 章では AI を導入・活用したいと考えている企業・自治体に求められる対応について提言したい。

2. AI のリスクマネジメントに関するフレームワーク

AI Risk Management Framework 1.0

本章では、2023 年 1 月にリリースされ、今後の環境変化に応じて適宜内容をアップデートすることが明言されているという理由から、AI Risk Management Framework 1.0（以下、「AI RMF」とする）について以下に要約・整理する。なお、AI RMF 本文は英文のため弊社にて日本語訳を行い、一般的に用いられている用語と齟齬のないように留意していること、今後正式に AI RMF の翻訳版・アップデート版が公表された場合には内容が変更される可能性があることについてご承知おき願いたい。

AI RMF は、以下に示す 3 つの観点について紹介されている。

・AI に求められる性能
・AI リスクマネジメントのための組織のありかた
・AI システムにおけるライフサイクルの考え

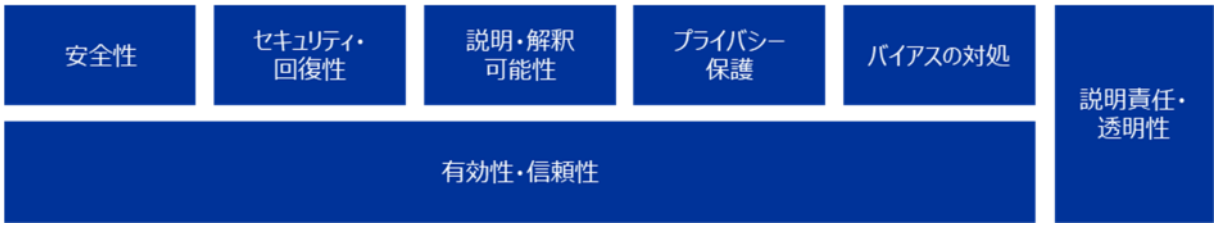
□AI に求められる性能

企業・自治体等で AI を開発・運用するためには、その AI がどれだけ信頼に値するものかという観点が最も重要であり、信頼に足る AI を構築するために管理されるべき項目として、安全性（Safe）、セキュリティ・回復性(Secure & Resilient)、説明・解釈可能性(Explainable & Interpretable)、プライバシー保護(Privacy-Enhanced)、バイアスの対処(Fair – With Harmful Bias Managed)、有効性・信頼性(Valid & Reliable)、説明責任・透明性(Accountable & Transparent)が取り上げられている（図表 4）。

また、これらの項目はトレードオフであり、すべてが高水準で適用されている AI というのは稀であるという説明もなされている。トレードオフの例として、AI を搭載した防犯カメラのように犯罪予防に係る AI を開発・実装することで安全性は増すが、監視の目が増えるためにプライバシーを少なからず損なうといったものが挙げられる。（本例は AI RMF に記述はなく、弊社にて追記）どのようなデータを入力/出力するのか、どのようなユーザー層に向けた AI システムなのかにより各項目の重要度が変わってくる。

■ 図表４ AI 構築において管理されるべき項目

有効性・信頼性は、上５つの性能に強く関与しているため、横長のボックスで示されており、説明責任・透明性は他のすべての項目に関連しているため、縦長のボックスで示されている。



大項目	小項目	概要	AI リスクマネジメントに資する取り組み
有効性・信頼性	有効性	計画された活動が実行され、結果が達成された程度	AI に業務を置き換えた際にどれだけのベネフィットが得られたかを確認する
	信頼性	与えられた条件で既定の期間中、要求された機能を果たすことができる性質	予測精度、汎化性能[24]を担保する
安全性	安全性	定義された条件下で人の生命・健康・財産・環境が危険に晒される状態にならないようにすること	優先度付けといった、リスク管理プロセスの実施を行う
セキュリティ・回復性	セキュリティ	不正なアクセス等の予期せぬインシデントに対する耐久性	要件定義の段階で、機密性に係るサービスレベルを定義する
	回復性	予期せぬインシデントがあっても機能を維持できる、もしくは安全に停止できる性能	要件定義の段階で、可用性に係るサービスレベルを定義する
説明・解釈可能性	説明可能性	AI システム内でどのように決定がなされたかについての指標	AI システムに用いたモデルの詳細について文書を残す
	解釈可能性	AI の出力結果に対し、「なぜ」この結果となったのかを説明できる根拠	AI 上で、機械学習に用いた各特徴量の寄与度を可視化するなど、結果出力の根拠を示す
プライバシー保護	プライバシー保護	推論によって個人情報や機密情報を特定しない/されないようにすること	プライバシーを侵害した場合の対応についての整理ならびにユーザーへの共有を行う
バイアスの対処	バイアスの対処	偏りのあるデータを AI に学習させることで、公平性のない偏った結果を算出してしまう事態を防ぐこと	学習前のデータ処理、学習後のデータ吟味を行い、入出力データにおける偏りを排除する取り組みを行う
説明責任・透明性	説明責任[25]	AI に使用しているデータセットの入手方法、使用方法や AI の出力結果等について適切に説明ができること	透明性に基づいて AI の危険性や原理を説明し、利用者からの合意を得る

	透明性	AI に関する情報を、ユーザーやステークホルダーが適切に理解できるかの指標	どのようなデータセットから学習したのか、どのように前処理したのかなど、AI の詳細情報を明示し、ステークホルダーに共有する

出典：AI RMF をもとに弊社作成

□AI リスクマネジメントのための組織のありかた ～AI RMF Core～

AI RMF では、信頼に値する AI を構築するために、データサイエンティストやプロダクトマネージャーなど、ステークホルダー同士がより密に関与するような実務の在り方が説明されている。具体的には統治（Govern）、位置づけ（Map）、評価（Measure）、管理（Manage）の４つに役割を区分し、それぞれについてどのようなマネジメント・連携が必要であるかが言及されている。以下図表５のとおり、Govern については他３要素の中心に位置しているが、これは情報提供・組織内文化の浸透を担う横断的な機能として設計されているためである。

■ 図表５ AI RMF Core

役割	求められる業務・要件例
統治	AI に関するプロセス、方針等が定まっているか AI 運用について、責任や権限の所在は決まっているか 組織内にリスクマネジメントの文化が醸成されているか
位置づけ	AI システム要件定義がなされているか AI の出力がどう使われるか AI が支援するタスクについて具体的に示されているか
評価	AI の性能が客観的に理解できる形で測定されているか AI ユーザーがシステムの成果についてフィードバックできる環境があるか
管理	AI システムの運用状況について定期的に評価・更新されているか サードパーティ製[26]の AI システム導入におけるリスク・利益が明文化されているか

出典：AI RMF（図）、AI RMF をもとに弊社作成（表）

□AI システムにおけるライフサイクルの考え

AI のリスクを特定・管理するためには、AI のライフサイクル全体を考慮した幅広い視点・ステークホルダーの把握が重要となってくる。図表 6 に示すようなアプリケーション・コンテキスト（Application Context）、データ・入力（Data and Input）、AI モデル（AI Model）、実装（Task and Output）、AI に係るユーザー・社会（People and Planet）からなる AI 設計のサイクルを AI のライフサイクルと表現している。

■ 図表 6 AI システムの構成ならびにライフサイクル

内側の 2 つの円が AI システムを構成している要素、一番外側の円が各要素における役割とそのライフサイクルを示している。

構成要素	役割	役割例	ステークホルダーの例
アプリケーション・コンテキスト	計画と設計	法律・規制等に配慮した AI システムのコンセプトや目的、仮定、要件等の明確化	エンドユーザー、プロダクトマネージャー
	運用と監視	AI システム実装後の運用・監視	エンドユーザー、システム運用部門
データ・入力	データ収集・処理	データセットの選定・分析・処理	データサイエンティスト、AI を対象とする分野の専門家
AI モデル	構築	AI モデルの訓練、作成	
	検証・調整	AI モデルの出力の検証や調整、解釈	
実装	AI システムの実用化	AI システムの実装、ユーザーの満足度の評価	SIer、ソフトウェアエンジニア、サードパーティ製品のサプライヤー
AI に係るユーザー・社会	AI 開発にあたっての情報共有	AI システムの開発における、リスク管理やガイダンス作成にあたっての背景情報の提供	エンドユーザー、研究者、AI システムに係る個人・団体

出典：AI RMF（図）、AI RMF をもとに弊社作成（表）

図表 6 に示すとおり、ライフサイクルを形成する各要素で関与する人材は異なってくる。AI RMF では、「AI リスクマネジメントの成功にはこれらの要素での連帯こそが重要である。AI システムに係るすべてのステークホルダーが知見や情報を共有することで、幅広い集団的視点をもち、より多くのリスクを特定できうる」と記述されている。

3. AI を導入・活用したいと考えている企業・自治体等に求められる対応

ここまで述べてきたように、AI の開発・導入・活用の際には、提供側、利用側双方にリスクが存在する。提供側がリスクマネジメントを考慮するのはもちろんのこと、利用側においても、AI RMF のライフサイクルで「全てのステークホルダーが協力してリスクを管理していく」と記述されているとおり、AI を原因とするリスクの全てが提供側の責任と考えず、利用側としてもリスクを管理していくことが望ましい。

本章では、今後ますます AI の導入・活用が見込まれる企業・自治体等において対応が必要となると推測される「利用者側の立場において、AI リスクをどのようにマネジメントすればよいのか」について、比較的簡易に取り組めるような事項を中心に述べていく。

（1）AI システム利用者側のリスクマネジメント体制の構築

AI RMF が他のマネジメントシステムとの整合[27]を意識して整理されていることから、まずは従来のリスクマネジメント体制（図表 7）に取り込み、リスク分類の一つとしてリスク特定～モニタリングを実践していくことが 1 つの手段だと考える。また、未然防止策にとどまらず、企業・自治体等としては AI システム起因での情報漏洩等のインシデントが発覚した際の対応や体制（危機管理体制）を事前に整理しておき、影響度・影響範囲を最小化できるようにすることが望ましい。

■ 図表 7 理想的なリスクマネジメントの進め方

出典：弊社HP より

□リスクマネジメント体制内における情報共有

AI RMF の「AI システムにおけるライフサイクルの考え」に記述されている通り、AI の利活用にはステークホルダーとの連携・利用者間での情報共有が重要である。利用者側として、どのような情報を利用者間若しくはステークホルダーと共有すべきかについて以下に示す。

［具体的な対策例］

- 既存のリスクマネジメント委員会等の組織体の議題として、AI に関するものも盛り込む。
- 既存の組織の下にタスクフォース/ワーキンググループのような形で、AI の利活用及びそこに伴うリスクについて対応を行うチームを形成する。
- ニュース等の外部情報等、AI に係る情報を継続的に収集し定期的に共有することで、組織内での AI に対する関心・正しい使い方等についての啓発に取り組む。
- 導入した AI システムについて、利用者数・アクセス数の推移を組織内で共有する。もしこれらの減少によりサービス継続による利益が見込めない場合は、AI システムの縮小/終了等について検討する。
- 導入した AI システムに対する使用感（要求・AI の精度等）について、組織内のアンケートフォーム等を通していつでも意見を把握できる体制を構え、定期的にベンダーへのフィードバックを行う。
- 利用者の AI システムへの入力ログを捉え、セキュリティ上問題がないかを定期的に報告する。
- 導入元の AI の規約事項を理解し、わかりやすい形に咀嚼して組織内共有することで、利用者に不適切な入力をしないよう促す。

（２）AI システムのリスクへの対策

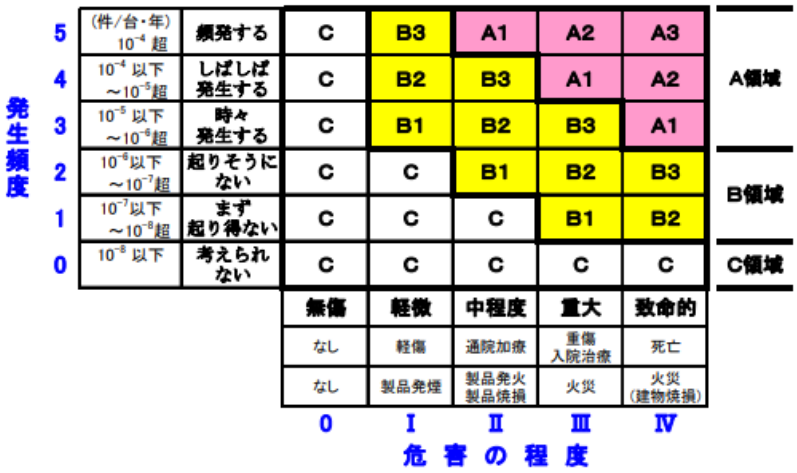
リスクマネジメントのなかで洗い出された優先度が高い AI リスクに対しては、利用者として対策を講じていく必要がある。前述のとおり、AI システムを原因とするリスクには様々な影響が想定されるが、「１．近年の AI を取り巻く状況について（３）AI 活用によるインシデントとその分類」で示したように影響分類で整理することができ、そのうえで AI リスクへの対策として従来のリスクへの対策方法が参考になる。なお、影響分類のうち「セキュリティ」については人間の悪意に基づいたサイバー攻撃が懸念されるため、本項では「悪用」に組み込む形で説明する。

□安全性

AI システムを活用する際の安全性・品質を担保することは供給側の責任の一つであるが、利用目的を超えた不適切な利用、管理不足による危害の発生に対しては、利用者側の責任が問われる可能性がある。例えば、工場の製造設備において AI ロボットを活用し、何らかの原因で従業員の負傷等が発生してしまった場合には、利用者である企業が従業員の負傷等に対して責任を負う可能性が考えられる。そのため、利用者側として導入しようとしている AI システムにどのようなリスクが潜んでいるかを可視化することがリスクマネジメントに資する。ここでは参考までに日本科学技術連盟[28]が開発した R-Map[29]手法を紹介する。

R-Map はリスクアセスメントのプロセスにおいてリスクを見積もり、評価するため、発生頻度（縦軸）と危害の程度（横軸）の観点から、A（許容不可）、B（危険/効用基準或いはコストを含めて低減策の実現性を考慮し最小限のリスクまで低減すべき領域）、C（許容可能）の３領域に分けてマッピングを行う手法である。（図表８）

■ 図表 8 R-Map



出典：経済産業省「リスクアセスメント・ハンドブック実務編 2011 年 6 月」より引用

［具体的な対策例］

- AI システム利用時のリスクについて、リスクアセスメントを実施する中で、R-Map に代表されるような既存のフレームワークを用いたマッピングを行い、適切にリスク評価をしたうえで、リスク低減対策を行う。
- AI システムについて、想定の範疇を超えた使い方をしないように心掛ける。
- AI システムを組成するハードの部分（カメラ・センサー等）について、定期点検を行う。
- AI システム等の使用について、適切なリスクアセスメント・KYT[30]活動を実施する。
- 従業員に対して、使用方法や禁止事項についての講習・教育を定期的に実施する。

□説明可能性

説明可能性が高い AI とは、AI が導出した出力に対して導出に至る過程や根拠まで示すことができる AI を指す。（近年ではこうした AI を「説明可能な AI（XAI）[31]」として表現されることが多い）一般的に、AI モデルは出力の導出過程が複雑ゆえ解釈するのは容易ではなく、中身がブラックボックスであるという説明をされることがしばしばある。しかし、利用者が AI システムを発注しようとしたときに、中身がどうなっているか分からないものを安易に採用できるだろうか。例えば、利用者の企業で開発した自動運転車が事故を起こしたとき、事故当時の AI の判断基準が分からなければ、企業の信頼を大きく失いかねない。このように AI システムの中身が分からないというのは利用者にとってのリスクとなりえる。利用者としては、AI の説明可能性に着目し、導入した AI システムの出力に対して導出過程・根拠を知ることができる体制を構えることが重要である。

[具体的な対策例]

- AI システム導入の基準のひとつに説明可能性を設ける。
- AI システムの提供者側に、出力の導出過程・根拠を確認できるような体制を構える。
- AI システムの利用規約や資料を読み、対象の AI システムでどこまで説明できるかを把握する。

□プライバシー

情報漏洩の原因として、機密情報の社外持ち出しや不適切管理、SNS の不適切な投稿などが考えられるが、AI の情報漏洩についても同様である。企業・自治体等の中で情報漏洩等はコンプライアンスの問題として

整理され、情報漏洩の対策としては従業員がコンプライアンス違反を行わないように、定期的な教育、定期的なアンケート・点検等を実施していくことが考えられる。

［具体的な対策例］

- 個人情報・機密データ AI システムに入力しない等の AI システム活用におけるガイドラインや基準を作成し、ルール化を図る。
- AI システム活用に関する情報漏洩問題をコンプライアンスの教育の内容として盛り込む。
- AI システム活用が組織内外のルールに違反していないかを定期的な点検項目に追加する。
- オープンなシステムではなく、組織内のみ（閉じられた空間での使用）で利用できるシステムとして AI を導入する。
- 万が一外部への情報の漏洩を確認した場合、発見者はすぐに運用部門に伝達し、運用部門はすぐに機能を停止・縮小できるようにしておく。

□公平性の欠如（バイアス）

利用者は AI システムによる出力が必ずしも適切で、正しい内容とは限らないと認識しておく必要がある。例えば、ChatGPT が公開している Terms of use（利用規約）[32]においても、誤った出力が発生する可能性について言及されている[33]。また、利用者による出力の活用においては、企業・自治体等や利用者に責任があることも留意する必要がある。AI の出力活用から危害等が生じた場合でも、AI システムを提供側の一方的な責任とならないよう規約[34]等で規定されることが想定されるため、活用の際は利用者自身に責任があることを留意しておく必要がある。企業・自治体等は従業員や職員に対してその認識を持たせることが対策の一つとなる。

また、AI 運用に係るメンバーを組織内だけで組むと、思考の偏りが発生する可能性がある。そうならないよう、社外からの人材確保や専門家からの情報収集など、人的ネットワークを確保することもバイアス排除に資すると考える。

［具体的な対策例］

- 外部講師・専門家からの協力を経て定期的に人権、著作権等にかかる研修を実施し、公平性が欠如した表現（バイアスのかかった表現）についての認識を共有する。
- ディープフェイク等、恣意的に悪意をもって AI システムに出力させるような操作を行わない。
- 出力が必ずしも適切ではないことを留意し、活用の際には企業・自治体等や従業員・職員個人に責任がある（利用に関して説明できる状態にする必要がある）ことを周知する。
- AI システムの利用において倫理的に不適切とされる出力を確認した場合は、すぐに運用部門に報告し利活用の方法を再検討する。

□悪用（＋セキュリティ）

生成型 AI の進歩に伴い、ディープフェイクを用いた悪用等が今後もあらゆる場面で発生するものと予想される。ディープフェイクの例として、なりすました人物の同僚や友人を騙す、組織内の人物の文書の書き方を模倣したビジネスメール詐欺[35]、画像生成技術を用いたフェイクニュースや詐欺広告等がある。これらはいずれも組織にとって大きな脅威となりうるため、こうした悪意に騙されることのないよう、より一層の注意を払う必要があるものと思料する。こうしたディープフェイクに対処するにはソーシャルエンジニアリング攻撃[36]への対策が参考になる。

[具体的な対策例]

- ディープフェイクに対する理解を組織内で共有する。
- ディープフェイクが疑われるようなケースについて整理し、コミュニケーションに関するルールを定める。（例えば、「パスワードや高額な金銭を求められた場合、直ぐに答えず本人に確認をとらせる」、「公衆電話や非通知など、番号が特定できない電話からの問い合わせには応じないようにさせる」、「受信したメールのドメインに注意させる」等）
- 組織内環境においてディープフェイク等の作成やテストを禁じる、あるいは制限を設ける。
- ディープフェイクに起因するあらゆる風評に対し、迅速に対応できるように備える。

□ヒューマンエラー

生成型 AI に機密情報を入力した際の AI に係るヒューマンエラーは、コンプライアンス意識の薄れ、AI に信頼を置きすぎること、導入された AI システムに対しての無知等が挙げられる。これらに対処するには、AI システムは何ができて何ができないのか、人間は AI システムに対して何をしてはいけないのかを正しく知ることが推奨される。

[具体的な対策例]

- 組織内でのコンプライアンス教育を徹底する。
- 導入する AI システムの規約・適用範囲を把握し、AI システムにできること、出来ないことを棚卸したのちに組織内に共有する。
- 導入の初期段階においては、複数名での指差し確認などを義務付ける。
- メンテナンス等、AI システムに対して人間は何をすべきなのかについてマニュアル等に整理する。

（3）企業・自治体等への提言

これまで、AI には様々なリスクがあることや、様々な国・組織が AI の利活用を目的としたリスクマネジメントの手法・フレームワークを検討していることとそのフレームワークの概要を本稿で紹介した。AI システムは、機械学習モデルや統計学、プログラミング等の技術のみならず、法律や倫理等が密に絡んで成り立っており、それゆえにリスクマネジメントの枠組みも複雑なものとなっている。これらの特徴を踏まえ、AI 利活用におけるリスクマネジメントについて、下記 3 点を提言したい。

- AI を利活用する際に、各人が AI リスクを理解することは当然重要であるが、個人で全容を把握することが困難であることも事実である。組織として、AI システムに係る社内外のステークホルダーが知見や情報を共有し合う場を設け、多角的な視点の下に AI リスクマネジメントを進めるべきである。
- 機密情報の厳重管理や機械の安全性確保等、従来のリスクマネジメント対策は、AI リスクマネジメントにおいても重要であるため、リスクマネジメント対策の内容を AI システムの特性にあわせて修正・強化するべきである。
- セキュリティや法律等、自組織内で完結することが困難なリスクも少なからず存在する。こうしたリスクに対処するために信頼できる外部専門家を選定し、問題があった際には相談できる関係性を構築しておくべきである。

4. おわりに

近年の AI 技術は加速度的に進歩し、その結果として様々な場所で話題に挙がるようになった。今後も利用者の増加が見込まれるため、AI 技術は様々な場面での利活用が期待できる。一方で、AI に係る様々なリスクが懸念されているため、AI の利活用に踏み切れない企業・自治体等も少なくないと推察する。リスクマネジメントを実践し、AI に係るリスクを把握することは、利活用を考えるにあたって非常に重要である。

本稿が、AI の利活用を検討される企業・自治体等において、少しでも気付きになれば幸いである。なお、本稿の記述において ChatGPT 等の文章生成 AI は用いていない。

[2023 年 5 月 30 日発行]

参考文献

「機械学習を解釈する技術　予測力と説明力を両立する実践テクニック」森下光之助　技術評論社（2021）

「ネット世論操作とデジタル影響工作」一田和樹，齋藤孝道，藤村厚夫，藤代裕之，笹原和俊，佐々木孝博，川口貴久，岩井博樹　原書房（2023）

「AI 社会の歩き方」江間有沙　化学同人（2019）

「AI 白書 2023」AI 白書編集委員会 編 株式会社角川アスキー総合研究所（2023）

---

[1] OpenAI 社により開発された会話型 AI サービス。 https://openai.com/blog/chatgpt

[2] 会話、ストーリー、画像、動画、音楽などの新しいコンテンツやアイデアを作成できる AI の一種。 https://aws.amazon.com/jp/blogs/news/announcing-new-tools-for-building-with-generative-ai-on-aws/

[3] G7 群馬高崎　デジタル・技術大臣会合，議論のテーマとして「責任ある AI と AI ガバナンスの推進」がある。 https://g7digital-tech-2023.go.jp/topics/topics_20230430.html

[4] 米国の NIST（National Institute of Standards and Technology）という機関により公表された AI 利活用に関するリスクマネジメントのフレームワーク。 https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF

[5] 文章を入力データに画像を出力する AI サービス。 https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F

[6] OpenAI 社が開発・提供している画像生成 AI ツール。 https://openai.com/product/dall-e-2

[7] Google 社が開発している会話型人工知能。2023 年 5 月より日本語に対応。 https://bard.google.com/

[8] OpenAI 社が開発・提供している会話型人工知能。 https://www.perplexity.ai/

[9] Amazon Web Services ブログ, “AWS で生成系 AI を使用した構築のための新ツールを発表”, https://aws.amazon.com/jp/blogs/news/announcing-new-tools-for-building-with-generative-ai-on-aws/

[10] Google, ”An important next step on our AI journey”, https://blog.google/technology/ai/bard-google-ai-search-updates/

[11] “AI アシスタントサービス「PX-GPT」をパナソニックグループ全社員へ拡大　国内約 9 万人が本格利用開始” https://news.panasonic.com/jp/press/jn230414-1

[12] “社内 AI チャット「Benesse GPT」をグループ社員 1.5 万人に向けに提供開始”, https://blog.benesse.ne.jp/bh/ja/news/management/2023/04/14_5969.html

[13] “三井化学、生成 AI と IBM Watson の融合による新規用途探索の高精度化と高速化の実用検証スタート”, https://jp.mitsuichemicals.com/jp/release/2023/2023_0412/index.htm

[14] 茨城県, ”日本初！自治体公認 Vtuber を AI 化　ChatGPT と AI 音声対話システムを連携した「AI 茨ひより」がニコニコ超会議 2023 に登場”, https://www.pref.ibaraki.jp/somu/hodo/hodo/pressrelease/hodohappyoushiryou/2203/documents/230420puromo.pdf

[15] Application Programming Interface の略称。ソフトウェアやアプリケーションの一部機能を外部に向けて公開することにより、第三者のソフトウェアでその機能を使えるようにするもの。

[16] 横須賀市のニュースリリースから抜粋。なお、このニュースリリースは ChatGPT で下案作成ののち職員により校正が行われた。出典：https://www.city.yokosuka.kanagawa.jp/0835/nagekomi/20230418_chatgpt.html

[17] 野村農林水産大臣記者会見概要における、「チャット GPT の一部業務での利用方針について」より抜粋。出典：https://www.maff.go.jp/j/press-conf/230418.html

[18] 日経サイエンス 2023 年 5 月号, ”特集　話す AI　描く AI”

[19] 機械学習アプローチを用い、事実にないような画像・動画などを生成すること。

[20] Partnership on AI　https://partnershiponai.org/

[21] AI INCIDENT DATABASE
https://incidentdatabase.ai/apps/discover/?display=details&is_incident_report=true&page=1&sortBy=relevance
[22] アメリカの IT 企業。コンピューター及びネットワークのセキュリティに関するソフトウェアを提供している。
出典：https://www.cyberhaven.com/
[23] Cyberhaven Labs が自社製品を使用する企業の 160 万人分労働者の ChatGPT の使用状況を分析したデータ。
出典：https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/
[24] 訓練データだけでなく、未知のデータに対しても適切に予測できる性能
[25] 何らかの不祥事やインシデントに対する釈明・謝罪をするという意味合いではない。
[26] ここでは他社製品の意。
[27] AI RMF 内にて、リスクマネジメント（ISO31000）が引用されている。特に、AI RMF Core は、リスクマネジメントの体制構築～運営を意識したものと推測される。
[28] 一般財団法人日本科学技術連盟。経営管理技術、「品質管理」を中心とする事業を展開している。
https://www.juse.or.jp/
[29] 日本科学技術連盟が開発した手法。経済産業省が公表しているリスクアセスメント・ハンドブック実務編において製品安全における代表的なリスクアセスメント手法として記載している。
https://www.meti.go.jp/product_safety/recall/risk_assessment_practice.pdf
[30] 危険（Kiken）予知（Yochi）訓練（Training）の略称。職場や作業に潜む危険性や有害性等の機器要因を発見し、解決する能力を高める手法。 参考：https://anzeninfo.mhlw.go.jp/yougo/yougo40_1.html
[31] AI モデルが導出した答えに対して人間が納得できる根拠を示す技術、またはそれに対する研究のことをさす。
出典：https://www.darpa.mil/program/explainable-artificial-intelligence
[32] OpenAI が公表している ChatGPT の利用規約。出典：https://openai.com/policies/terms-of-use
[33] ChatGPT Term3.Content (d) Accuracy
[34] ChatGPT の利用規約においても、Term7. Indemnification; Disclaimer of Warranties; Limitations on Liability で、ChatGPT の使用に起因または関連する損失、費用等についての免責事項が記載されている。
[35] 自社または関連会社の経営者・取引先になりすまして金銭を騙し取ることを目的としたサイバー攻撃。
出典: https://www.ipa.go.jp/security/bec/index.html
[36] パスワード等の重要な情報を、情報通信技術を使わずに盗み出すサイバー攻撃。その多くは人間の心理的な隙や行動のミスにつけ込むもの。手法としては企業の上層部に成りすまして高圧的にパスワードを要求する、公共の場での覗き見、ゴミ箱から機密情報が書かれた書類を漁る、などがある。
出典： https://www.soumu.go.jp/main_sosiki/cybersecurity/kokumin/business/business_staff_12.html

To Be a Good Company
東京海上ディーアール株式会社

ビジネスリスク本部　研究員　加藤 直人
専門分野：製造業における営業データ分析、海上輸送コンテナ積載シミュレーションシステムの設計開発等に従事。統計学、データ分析を活用したリスクマネジメント体制構築、製造業に対する事業継続計画策定等を支援。
ビジネスリスク本部　主任研究員　木村 圭佑
専門分野：製造業における製品開発・品質保証に従事。製造業、建設業、小売業、金融・保険業、広告業に対する事業継続計画策定、災害対応訓練、リスクマネジメント体制構築等を支援。
〒100-0004 東京都千代田区大手町 1-5-1 大手町ファーストスクエア ウエストタワー23F
Tel. 03-5288-6594　Fax. 03-5288-6626 https://www.tokio-dr.jp/