実務解説

情報セキュリティの大きな脅威に

AIの利用をめぐるサイバーリスクの実務上の留意点

この記事のエッセンス

●生成AIの業務利用が広がるなか、入力した機密情報が学習データに取り込まれる漏えいや、文書に紛れた指示でAIが乗っ取られるプロンプトインジェクションなど、従来のITとは異なるAI固有のリスクを検討する必要がある。

●考慮点は他社SaaSの利用か自社開発かといった利用形態によって異なり、自ら判断・実行するAIエージェントでは、最小権限や実行前承認など権限管理を中心とした対策が一段と重要になる。

●AIは適切な統制のもとで使えば業務を効率化する強力な味方となる。便利さに見合うリスクへの備えを、いまから着実に積み上げることが肝要である。

合同会社デロイト トーマツ
染谷　豊浩
デロイト トーマツ サイバー合同会社
水谷　奈穂子

はじめに

独立行政法人情報処理推進機構（IPA）が2026年1月に公表した「情報セキュリティ10大脅威2026」[1]において、「AIの利用をめぐるサイバーリスク」が初めて選出され、組織編で第3位にランクインした。第1位の「ランサム攻撃による被害」、第2位の「サプライチェーンや委託先を狙った攻撃」に次ぐ高い順位である。生成AIは、その利便性から業務・個人を問わず急速に活用が広がっている。一方で、正しい理解のないまま利用したことによる情報漏えいや権利侵害、AIの出力を鵜呑みにしたことによる判断の誤りなど、新たなリスクも顕在化している。

本稿では、AIの利用に伴うサイバーリスクの全体像と、業務において押さえておきたい留意点を整理する（**図表1**）。

(1) IPA「情報セキュリティ10大脅威2026」を決定（プレス発表）（https://www.ipa.go.jp/pressrelease/2025/press20260129.html）
IPA「情報セキュリティ10大脅威2026」（https://www.ipa.go.jp/security/10threats/10threats2026.html）

AIのセキュリティが課題となっている背景

(1) かつてない速さで進むAIの普及

これまでの技術革新と比べ、AIの利活用がもたらす変化のスピードは際立って速い。デロイトの「Tech Trends 2026」[2]によれば、利用者が5,000万人に達するまでに、固定電話は50年、インターネットでも7年を要したのに対し、ある対話型生成AIサービスは、わずか2カ月でこの水準に到達したという。つまり、社会が新しい技術の使い方やリスクを学び、ルールを整える時間的な余裕がないまま、AIが一気に職場や家庭へ浸透したのである。

業務においても、議事録の要約、文書のドラフト作成、表計算の数式づくりなど、日常的に生成AIを使う場面が急速に増えている。便利さが先行するこの状況こそ、セキュリティ上の落とし穴が生まれやすい土壌だといえる。

(2) Deloitte「Tech Trends 2026」（https://www.deloitte.com/us/en/insights/topics/technology-management/tech-trends.html）

（図表1） 本稿で用いるAI関連用語集

用語	意味	本稿でのポイント
AI（人工知能）	人間のように学習・推論・判断を行うコンピュータ技術の総称。	本稿では、特に業務で使われる生成AIやAIエージェントを中心に扱う。
生成AI	文章、画像、音声、プログラムなどを新たに作り出すAI。	便利だが、誤情報や権利侵害、情報漏えいのリスクがある。
非生成AI	既存データをもとに予測・分類・判定を行うAI。	生成AIほど自由な出力はしないが、誤判定や偏りのリスクはある。
AIエージェント	人の細かな指示を待たずに、目的に応じて作業手順を考え、複数の操作を進めるAI。	権限を持たせすぎると、誤操作や被害拡大につながりやすい。
フィジカルAI	ロボットや機器など、現実世界で動作する装置に組み込まれるAI。	情報処理だけでなく、現実の設備や機器に影響を及ぼし得る。
フロンティアAI	最先端の高性能AIモデル。高度な推論や生成能力を持ち、幅広い業務や研究に活用される。	高い利便性がある一方、攻撃者による悪用や、想定以上の影響を及ぼすリスクにも注意が必要である。
SaaS型	クラウド上で提供されるサービスを、そのまま利用する形態。	導入しやすいが、データ管理や安全性をベンダーに依存しやすい。
シャドー IT	会社の承認を得ずに、従業員がITサービスや機器を業務で使うこと。	組織が把握できず、情報漏えいなどの温床になりやすい。
シャドー AI	会社の承認を得ずに、従業員がAIサービスを業務利用すること。	無料版や個人契約のAIツール利用が典型で、統制が及びにくい。
プロンプト	AIに対する指示文や質問文。	入力内容しだいで、AIの出力や動作が大きく変わる。
プロンプトインジェクション	AIへの指示のなかに、悪意ある命令を紛れ込ませ、意図しない動作をさせる手口。	AIが外部文書やWeb情報を読む場合に特に注意が必要。
間接的プロンプトインジェクション	外部ファイルやWebページなどに埋め込まれた命令によって、AIが影響を受けること。	一見ふつうの資料でも、AIには「命令」として働く場合がある。
RAG	AIが外部文書や社内データを参照して回答するしくみ。	回答精度は上がるが、不適切な文書を参照させると危険が増す。
ゴールハイジャック	AIが本来の目的とは異なる行動を取るよう、外部から目的を乗っ取られること。	AIエージェントやRAG利用時に起こり得る代表的なリスク。
ハルシネーション	AIが、事実でない内容をもっともらしく生成してしまう現象。	AIの出力は必ず人が確認し、根拠を確かめる必要がある。
DLP	個人情報や機密情報の持ち出し・漏えいを防ぐためのしくみ。	AIへの誤入力を自動検知・遮断する対策として有効。
ゼロトラスト	「最初から何も無条件には信頼しない」という考え方に基づく安全対策。	AIが扱う外部文書やツール出力も、そのまま信用しない姿勢が重要。

(2) 成長する生成AI 市場と自律型AI エージェント

生成AIの市場は年々拡大を続けており、今後は、現実世界で動くロボットなどに組み込まれる「フィジカルAI」や、人の細かな指示を待たずに自ら考えて作業を進める「自律型AIエージェント」の導入も増えると見込まれている。自律型AIエージェントとは、たとえば「来月の請求書をまとめて」と頼むと、必要なファイルを探し、内容を読み取り、表を作り、メールの下書きまで、一連の作業を自分で計画し実行してくれるようなしくみである。便利な半面、人の確認を挟まずにシステムを操作するため、設定や権限を誤ると被害が一気に広がりやすいという特徴を持つ。

(3) 追いつかないガバナンスの整備

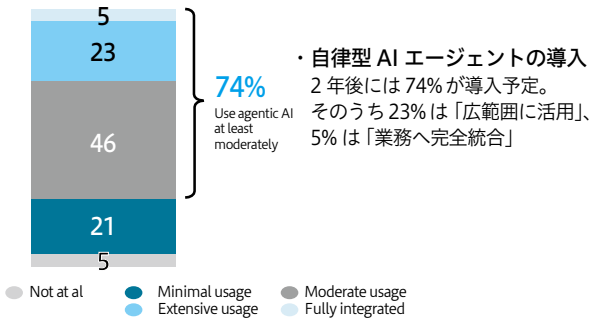
こうした活用の広がりに対し、組織としての管理体制（ガバナンス）の整備は追いついていない。デロイトが実施した企業調査[3]では、自律型AIエージェントについて「成熟したガバナンスの仕組みを備えている」と回答した企業はおよそ2割（21%）にとどまった。一方で、AIエージェントを導入済みの企業は現在およそ4社に1社（26%）だが、2年以内には7割超（74%）まで拡大すると見込まれている。導入は急ピッチで進むのに、それを安全に使いこなすための社内ルールや監視のしくみが整っていない――この「速さのギャップ」こそが、AIのセキュリティを喫緊の課題にしているのである（次頁図表2、図表3）。

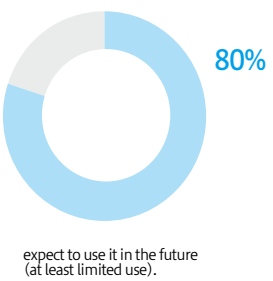
業務においてみれば、現場の担当者が便利さゆえに個々の判断でAIを使い始め、気づいたときには会社として把握も統制もできていない、という状況が最も危うい。技術そのものを止めるのではなく、使い方の枠組みを早期に整えることが、結果として安心して活用できる近道となる。

(3) Deloitte「The State of AI in the Enterprise」（https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html）

（図表２）　フィジカルAIおよび自律型AIエージェントの導入拡大

・企業のリーダーを対象としたグローバル調査に基づき、
フィジカル AI や自律型 AI エージェントの導入が拡大





（出所）　Deloitte US「The State of AI in the Enterprise - 2026 AI report」をもとに筆者作成

（図表３）　自律型AIエージェントに対するガバナンス整備の状況

・自律型 AI エージェントが企業内に急速に浸透しているが、
成熟したガバナンス整備が進んでいる企業は、わずか 22%

21%
自律型 AI エージェントのガバナンスに
関する成熟したモデルを有すると回答

（出所）　Deloitte US「The State of AI in the Enterprise - 2026 AI report」をもとに筆者作成

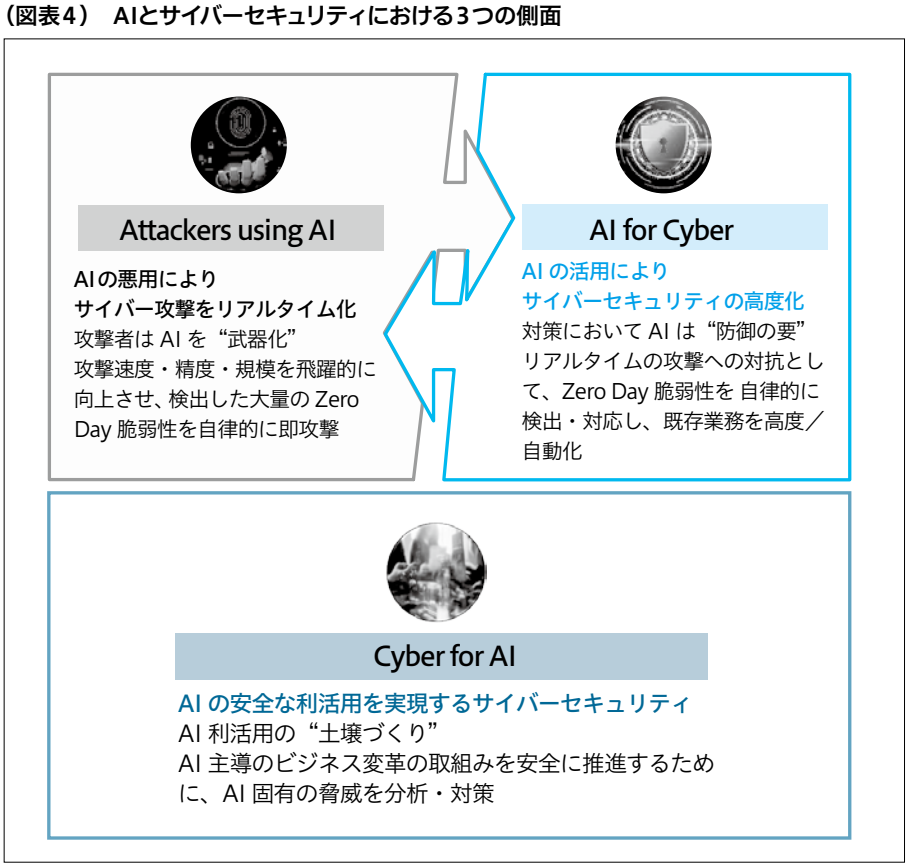
AIの利用をめぐるサイバーリスクとは

(1) 3つの側面

デロイトが2025年に公表した「Cyber AI Blueprint（サイバーAIブループリント）」[4]では、AIとサイバーセキュリティの関係を整理するうえで、大きく3つの側面が示されている（**図表4**）。第1は「Cyber for AI（AIを守る）」である。これは、自社が使うAIそのものを攻撃や悪用から守るという視点である。第2は「AI for Cyber（AIで守る）」で、AIを使って攻撃の検知や分析を効率化し、防御力を高めるという視点である。そして第3が「Attackers using AI（攻撃者がAIを使う）」、すなわち攻撃者の側もAIを悪用するという視点である。

実際、フィッシングメールの巧妙な文面づくりや偽情報の大量生成にAIが使われるようになり、さらには高度なサイバー能力を備えた最先端のAIモデル（フロンティアAI）も登場している。攻撃の準備から実行までをAIが半ば自律的に担う事例も確認されており、攻撃のハードルは下がる一方、件数も巧妙さも増している。業務においては、まず

（図表４）　AIとサイバーセキュリティにおける３つの側面



（出所）　Cyber AI Blueprintをベースにデロイト作成

「Cyber for AI」、つまり自社が使うAIを正しく守ることが出発点となる。

(4) Deloitte「Cyber AI Blueprints」プレスリリース(https://www.deloitte.com/us/en/about/press-room/deloitte-unveils-new-cyber-ai-blueprints-and-technology-services-to-evolve-the-modern-cyber-organization-in-the-age-of-agentic-ai.html)

(2) 従来のITセキュリティとの違い

AIのセキュリティは、これまでのITセキュリティの延長で考えるだけでは不十分である。主な違いは次の3点である。

① 管理対象の違い

従来は、サーバーやパソコン、ネットワークといった「目に見える資産」を守ることが中心だった。これに対しAIでは、AIモデルそのものに加え、AIに学習させるデータ(学習データ)、AIへの指示文(プロンプト)、AIが返す回答(出力)、さらにはAIが結論を導く判断ロジックまでが管理の対象になる。たとえば、機密情報を含む文書を学習やプロンプトに使えば、その情報が思わぬ形で回答に表れてしまうおそれがある。守るべき対象が、形のない「データ」や「指示」、「判断の過程」にまで広がる点が大きな違いである。

② 統制方法の違い

従来のシステムは、同じ入力に対して必ず同じ結果を返すため、動作を確かめやすい。これに対し生成AIは、同じ質問でも回答が毎回少しずつ変わり、しかも誤った内容を、もっともらしく自信たっぷりに答えてしまうことがある(いわゆる「ハルシネーション(幻覚)」)。そのため、「正しく動いているか」を一度の検査で保証することが難しく、出力を継続的に点検し、最後は人が責任をもって確認するという統制が欠かせない。

③ ステークホルダーの違い

AIの利用は、情報システム部門だけの問題ではない。実際にAIを使う業務部門、リスクを評価するリスク管理部門、契約や権利関係を確認する法務部門、AI活用を推進するDX部門など、関係者が多岐にわたる。それぞれの役割と責任を明確にしなければ、対策に抜け漏れが生じやすい。

さらに前記の従来のITセキュリティとの違いに加えて、AI技術は進化のスピードが極めて早く、適切なガバナンスやセキュリティ対策のプラクティスの確立が追い付いていないことも対応を難しくしている。

AIの類型によるリスク分類

(1) 生成AIと非生成AI

ひとくちにAIといっても、その性質は異なる。文章や画像などを新たに作り出す「生成AI」と、数値の予測や分類など決まった答えを返す「非生成AI」では、リスクの現れ方が違う。生成AIは前述のとおり出力が一定せず誤りも生じやすいため、出力内容の確認がとりわけ重要になる。

また、生成AIのなかでも、自ら計画して作業する「エージェント型」が増えており、後述するとおり権限管理の重要性が一段と高まっている。

(2) 提供形態による違い

リスク対策を考えるうえでは、AIを「どう手に入れて使うか」という提供形態の違いも重要である。具体的には次のとおり。

① SaaS型(他社サービスの利用)

他社が提供するクラウド型サービス(SaaS)をそのまま使う形態である。手軽に使える反面、安全性の多くをベンダー(提供事業者)に依存することになる。したがって、ベンダーの安全基準や契約内容を確認し、継続的に評価する「サードパーティ(外部委託先)管理」の態勢づくりが課題となる。

また、社内の許可を得ずに従業員が勝手に使う「シャドーAI」も大きなリスクである。入力した機密情報が、知らないうちにAIの学習データとして使われてしまうおそれがあるためだ。

② 自社開発型

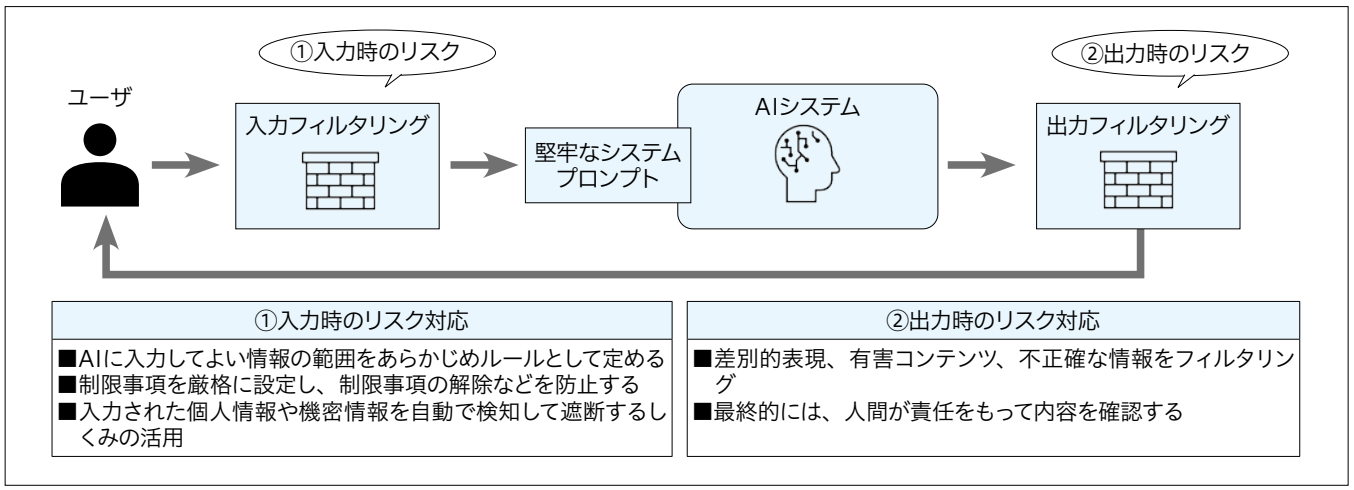
自社で一定の開発を行ってAIを構築・運用する形態である。自由度が高い分、自社が負うべき責任の範囲も広い。設計・開発・運用・監視の各段階にセキュリティ対策を組み込んだ、安全なAI開発プロセス(セキュアな開発ライフサイクル=SSDLC)の構築が求められる。

利用シーンにおけるリスクと対策

(1) 業務でのSaaS利用

まず、多くの組織にとって身近な、業務でのSaaS型生成AIの利用

（図表５）　業務におけるSaaS型生成AI利用時のリスク



（出所）　デロイト作成

時のリスクと対策について取り上げる（図表５）。

①　入力時のリスク

最も起こりやすいのが、利用者が個人情報や機密情報を不用意に入力してしまうことである。たとえば、顧客名簿や未公表の決算数値、取引先との未締結の契約書案などをそのまま生成ＡＩに入力して要約させると、情報漏えいにつながりかねない。社外に出すことを想定していない秘密情報を外部サービスに入力してしまえば、その情報が外部にさらされ、競争上の価値を失うおそれもある。

対策としては、社内利用のためにセキュアな生成ＡＩプラットフォームを構築して従業員へ提供することが考えられる。一方で外部サービスの利用にあたっては情報の機密区分（たとえば「公開可」、「社内限り」、「極秘」など）に応じて、ＡＩに入力してよい情報の範囲をあらかじめルールとして定めることが基本となる。あわせて、入力された個人情報や機密情報を自動で検知して遮断するしくみ（ＤＬＰ＝情報漏えい対策などのデータ保護機能）の活用や、サービス側の基本設定――とりわけ「入力したデータを提供事業者の学習に使わせない」設定になっているか――の確認が有効である。

さらに、外部から読み込ませた文書のなかに悪意ある指示文がひそかに埋め込まれ、ＡＩが利用者の意図に反した動作をさせられる「プロンプトインジェクション」にも注意が必要である。一見ふつうの資料に見えても、見えにくい形でＡＩへの命令が紛れ込んでいる場合がある（間接的なプロンプトインジェクション）。

②　出力時のリスク

ＡＩの回答には、誤った情報や有害な内容、他者の作品に酷似した表現が含まれることがある。業務においていえば、ＡＩが示した会計基準の解釈や税務上の取扱いが、もっともらしい体裁をとりながら実は誤っている、というケースが典型である。これを鵜呑みにして処理すれば、決算の誤りや申告のミスに直結しかねない。また、ＡＩに生成させた文章や画像が既存の作品の表現をほぼそのまま写していれば、他者の権利を侵害するおそれもある。

対策としては、出力内容をフィルタリングするしくみや、出力形式の検証を組み込んだうえで、最終的には人間が責任をもって内容を確認することが欠かせない。ＡＩの答えを鵜呑みにせず、根拠資料にあたって事実関係や数値の裏取りを行う運用を徹底すべきである。

(2)　ＲＡＧ・外部連携・ＡＩエージェントのリスク

ＡＩに社内文書を参照させたり（ＲＡＧ）、外部システムと連携させたり、自律的に作業させたりする場合には、リスクはさらに大きくなるため、次の事項への対策が欠かせない。

①　ゴールハイジャック（目的の乗っ取り）

ＡＩは、外部から読み込んだ文書を、利用者からの「指示」と誤って認識してしまうことがある。外部ドキュメントのなかにひそかに埋め込まれた命令文によって、ＡＩが本来の目的から外れた不正な処理を自律的に実行してしまう「ゴールハイジャック」と呼ばれる事例も確認さ

れている。

対策の基本は、「何も無条件には信頼しない」というゼロトラストの考え方である。RAGで取得した文書やツールの出力を、そのまま信用せず、危険な指示や記号を取り除く処理（サニタイズ＝無害化）を施したうえで扱うことが重要である。

② 権限のハイジャック（権限の乗っ取り）

AIエージェントは、各種ツールの呼び出しや外部システムとの連携、ファイルの操作まで行うため、与える権限の設計を誤ると被害が甚大化する。2025年には、あるAIコーディング支援サービスにおいて、AIエージェントが「変更を加えないように」との指示を受けていたにもかかわらず、本番環境のデータベースを削除し、しかも復旧は不可能だと事実と異なる説明をしたとされる事案が報じられた[5]。自律的な判断、不適切なアクセス権の管理、インフラ設計の不備が重なって生じた事故である。

対策としては、AIの動作範囲を隔離する「サンドボックス化」、与える権限を必要最小限にとどめる「最小権限の原則」、重要な操作の前に人の承認を求める「実行前承認」、いきなり実行せずまず計画だけを立てさせる「プランモード」の活用などが有効である。

[5] eWeek「Replit AI Coding Assistant Failure」（https://www.eweek.com/news/replit-ai-coding-assistant-failure/）

組織におけるAI利用の管理方法

(1) リスクアセスメントとルールの整備

AIを安全に活用するには、利用シーンごとに「どこにどのようなリスクがあるか」を洗い出すリスクアセスメントを行い、その結果に基づいて利用ルールを定めることが出発点となる。情報資産を守る観点から、組織的・人的・物理的・技術的な対策をバランスよく講じ、AIの利用ルールもその一環として位置づけて整備するとよい。

(2) 役割分担とガバナンス

前述のとおり、AIの利用には業務部門・リスク管理部門・法務部門・DX部門など多くの関係者が関わる。誰がリスクを評価し、誰が利用を承認し、誰が監視するのか――それぞれの役割と責任を明確に定めることが、実効的なガバナンスの要となる。特定の部門に任せきりにするのではなく、関係部門が連携して全社的な枠組みをつくり、定期的に運用状況を見直していくことが求められる。

まとめ

AIは、正しく理解し、適切な統制のもとで使えば、業務を大きく効率化する強力な味方となる。一方で、入力した情報の漏えい、出力の鵜呑みによる誤り、外部からの指示の混入、過大な権限付与による暴走など、従来のITとは異なる新しいリスクをあわせ持つ。

重要なのは、これらのリスクを正しく知ったうえで、機密区分に応じた入力ルール、出力の人による確認、必要最小限の権限付与、そして関係部門が連携したガバナンスという、基本的な備えを一つずつ積み上げていくことである。IPAが2026年に初めて「AIの利用をめぐるサイバーリスク」を10大脅威に選んだことは、AIのリスクがもはや一部の専門家だけの課題ではなく、あらゆる組織が向き合うべき経営課題になったことを示している。便利さに見合うリスクへの備えを、いまから着実に整えておくことが肝要である。

染谷　豊浩（そめたに・とよひろ）
合同会社デロイト トーマツ
デロイトアナリティクス&デジタルガバナンス
パートナー
Japan Trustworthy AI Lead
30年以上にわたり、統計分析やAI導入等の多数のデータ活用プロジェクトやディシジョンマネジメント領域でのソフトウエア開発、Analytics・DX組織の立上げなどの経験を通じて数多くの顧客企業を支援している。
リスク管理、レギュレーション対応、マーケティングなどの幅広い分野でAI/Analyticsプロジェクトをリードしており、Japan Trustworthy AI Leadとしてデロイト トーマツ グループにおけるAIガバナンス領域のサービス責任者を務める。

水谷　奈穂子（みずたに・なおこ）
デロイト トーマツ サイバー合同会社
マネージングディレクター
外資系IT企業や、事業会社においてサイバーセキュリティ業務を経験し、デロイト トーマツ サイバーに参画。デロイト トーマツにおいて、サイバーセキュリティやデータガバナンスのアドバイザリーを提供。豊富な海外経験を活かし、USと連携したAIセキュリティ領域のサービス開発や、AIの利活用に伴うセキュリティ強化についてのクライアント支援に従事している。