AI利活用におけるセキュリティ対策と法的責任

～AI事業者ガイドラインを踏まえて～

2026年２月14日

TMI総合法律事務所
弁護士　柴野　相雄

自己紹介　弁護士　柴野　相雄

- 1998年　慶應義塾大学法学部法律学科卒業、**2002年　TMI総合法律事務所　入所**
- 2010年　ワシントン大学ロースクール（LL.M., Intellectual Property Law and Policy コース）卒業、2010年9月～2011年5月　サンフランシスコのモルガン・ルイス＆バッキアス LLP勤務
- 2016年　慶應義塾大学法科大学院　非常勤教員就任（知的財産法務ワークショップ・プログラム）、2018年　一橋大学大学院　法学研究科　ビジネスロー専攻　非常勤講師（デジタル時代の著作権法）（隔年）、2019年　ISO/PC 317 (Consumer protection: Privacy by design for consumer goods and services) 国内審議委員会　委員就任、2022年2月　東京大学　未来ビジョン研究センター　客員研究員就任（～2025年1月31日）、2022年6月　一般社団法人　外国映画輸入配給協会　理事就任、デジタル庁技術検討会議　ガバメントソリューションサービス　タスクフォース　専門委員就任、2023年1月　慶應義塾大学大学院　政策・メディア研究科　特任教授　就任、2025年1月　農林水産省　優良品種の管理・活用のあり方等に関する検討会分科会委員　就任。

- 知的財産法、情報の保護に関する法分野、電子商取引に関する法分野を専門としており、IT、インターネットビジネス、エンタテインメント、広告、メディアに関する裁判、法律相談等を多く扱う。
- 近時の主な著書として、「IT・インターネットの法律相談［改訂版］」（2020年　青林書院）、「AIDCプラットフォームにおけるデータ提供契約に関する報告書」（2022年　一般社団法人AIデータ活用コンソーシアム）、「ヘルスケアビジネスの法律相談」（2022年　青林書院）、「個人情報管理ハンドブック（第5版）」（2023年　商事法務）等がある。
- 連絡先　　E-mail：tshibano@tmi.gr.jp

本日の内容

I. AI事業者ガイドラインとAIガバナンス

II. セキュリティの確保

1. セキュリティ確保について
2. セキュリティ確保に関する記載
3. AIによる便益/リスク
4. AI開発者向け　セキュリティ確保のポイント
5. AI提供者向け　セキュリティ確保のポイント

6-1. AI利用者向け　セキュリティ確保のポイント

6-2. AI利用者向け　セキュリティ確保の具体的な手法

III. AIのセキュリティ確保のための技術的対策に係るガイドライン（案）

IV. システム開発におけるセキュリティ関連裁判例

V. 生成AIサービスにおけるセキュリティ対策と法的責任

Ⅰ.AI事業者ガイドラインとAIガバナンス

「AI事業者ガイドライン」の構成と対象者

- AI事業者ガイドライン（1.1版）（令和7年3月28日）（総務省、経済産業省）
- https://www.soumu.go.jp/main_content/001002576.pdf（本編）
- 「AI事業者ガイドライン」は、本編と別添で構成され、「AIの安全安心な活用が促進されるよう、我が国における**AIガバナンス**の統一的な指針を示す」ものであり、リスクベースアプローチ（※）にもとづいて作成。
  - （※）「予め事前に当該利用分野における利用形態に伴って生じうるリスクの大きさ（危害の大きさ及びその蓋然性）を把握したうえで、その対策の程度をリスクの大きさに対応させる」アプローチ（本編3頁）

- 本編は、第2部「AIにより目指すべき社会及び各主体が取り組む事項」において、AIにより目指す社会としての「基本理念」を掲げ、その実現に向けて各主体が取り組むことが期待される「原則」とともに、そこから導き出される「共通の指針」を記載しています。

➢ **「共通の指針」は、①人間中心、②安全性、③公平性、④プライバシー保護、⑤セキュリティ確保、⑥透明性、⑦アカウンタビリティ、⑧教育・リテラシー、⑨公正競争確保、⑩イノベーションという10の指針から構成されています。**

- 「AI事業者ガイドライン」の対象者は「様々な事業活動においてAIの開発・提供・利用を担う全ての者（政府・自治体等の公的機関を含む）」とされています。
- 特に本編では、その対象者を**「AI開発者」「AI提供者」「AI利用者」**の3つに大別し、それらが念頭に置くべき基本理念（= why）、および、その理念を踏まえてAIに関して行うべき取組みの指針（= what）を示しています。

AIガバナンス　AIによる便益

- **AI ガバナンス**　（1.1版本編本文　10頁）
- AI の利活用によって生じるリスクをステークホルダーにとって受容可能な水準で管理しつつ、そこからもたらされる正のインパクト（便益）を最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計並びに運用。

- **AI による便益主な記載内容**
- （1.1版別添概要6頁）

- 便益を享受する最終利用者に焦点を当て、業種や業務ごとにAIによる便益を整理。

- https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_4.pdf

	開発	マーケティング	販売	物流・流通	顧客対応	法務	ファイナンス	人事
従来から存在する便益の例	コード検証、ドキュメント作成の自動化	広告用メールの自動配信	受注後の対応メール等の自動発信	需要予測に基づく生産・在庫数最適化	チャットボットによる自動対応	翻訳	財務諸表の自動作成	給与計算等の自動化
生成AIで更に向上	類似コード・データの抽出・検証	データに基づいたパーソナライゼーション広告	チャネル別、ニーズ別の売上予測	配送ルート最適化	過去の問合せ内容に基づいたFAQ作成	法務文章のレビュー	過去実績にもとづいた将来予測、不正検知	職務経歴書等に基づいた人材需要マッチング
生成AI特有の便益の例	学習データの生成、コーディングアシスタント、新製品のブレインストーミング	販売促進(マーケティング素材・キャッチコピー等)の自動作成	営業トークスクリプトの自動作成	物流条件交渉のアシスタント	対応内容の自動生成、要約	規定に基づいた契約書ドラフトの自動生成	文脈を踏まえた上での社内問合せ対応	文脈を踏まえた上での人事面接の対応

AIガバナンス　AIによるリスク

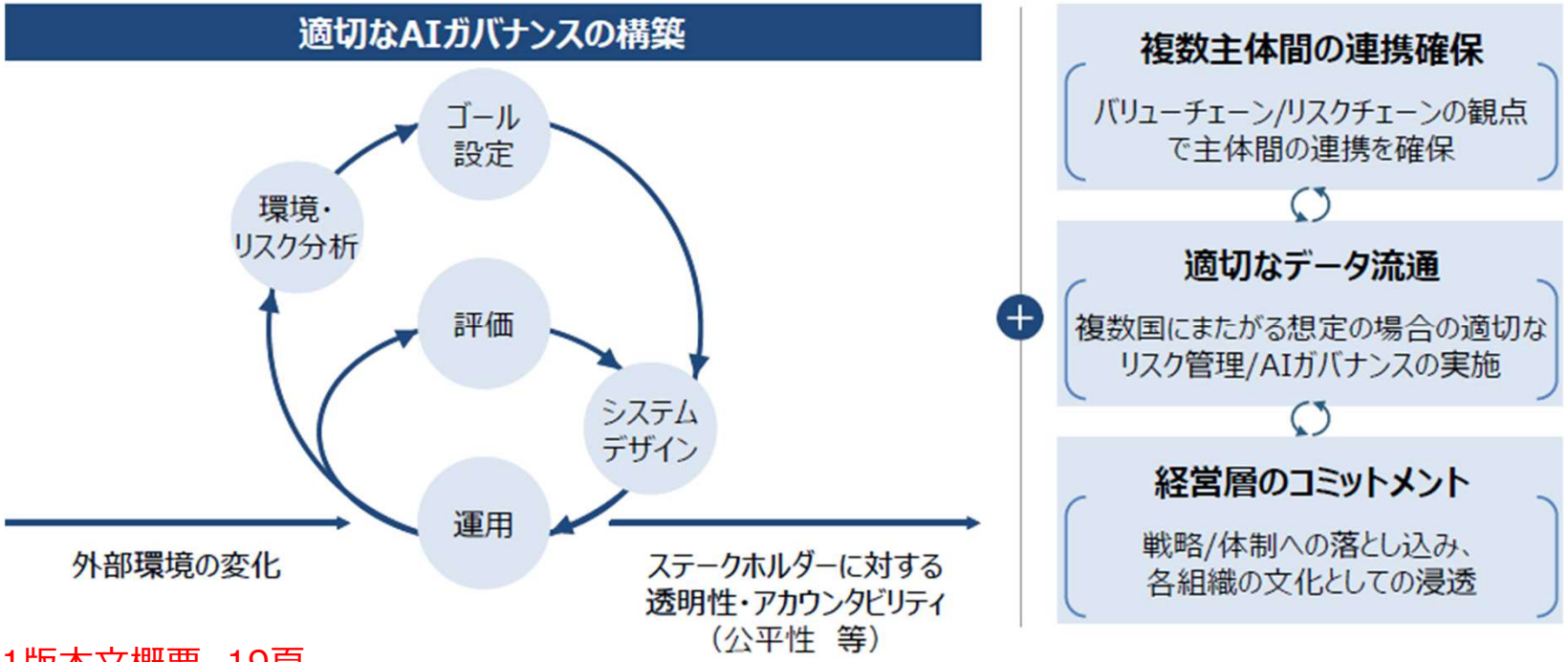
- **AI によるリスク**
- **主な記載内容**
- （1.1版別添概要7頁）

- AIによるリスクを、事業者ができる限り網羅的に把握し対策を検討できるよう、体系的に整理。
- https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_4.pdf

大分類	中分類	リスク例
技術的リスク（＝主にAIシステム特有のもの）	学習及び入力段階のリスク	データ汚染攻撃等のAIシステムへの攻撃
	出力段階のリスク	バイアスのある出力、差別的出力、一貫性のない出力等
		ハルシネーション等による誤った出力
	事後対応段階のリスク	ブラックボックス化、判断に関する説明の不足
社会的リスク（＝既存のリスクがAIにおいても発生又はAIによって増幅するもの）	倫理・法に関するリスク	個人情報の不適切な取扱い
		生命等に関わる事故の発生
		トリアージにおける差別
		過度な依存
		悪用
	経済活動に関するリスク	知的財産権等の侵害
		金銭的損失
		機密情報の流出
		労働者の失業
		データや利益の集中
		資格等の侵害
	情報空間に関するリスク	偽・誤情報等の流通・拡散
		民主主義への悪影響
		フィルターバブル及びエコーチェンバー現象
		多様性・包摂性の喪失
		バイアス等の再生成
	環境に関するリスク	エネルギー使用量及び環境の負荷

AIガバナンスの構築　経営において求められること

- AIを安全安心に活用していくために、経営層のリーダーシップのもと、下記に留意しながら適切なAIガバナンスを構築することで、リスクをマネジメントしていくことが重要となります
  - 複数主体に跨る論点について、バリューチェーン/リスクチェーンの観点で主体間の連携確保
  - 上記が複数国にわたる場合、データの自由な越境移転の確保のための適切なAIガバナンスの検討
  - 経営層のコミットメントによる、各組織の戦略や企業体制への落とし込み/文化としての浸透



- 1.1版本文概要　19頁
- https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_2.pdf

AIガバナンスの構築において留意する観点としての行動目標（別添本文23頁）

- 常に変化する環境及びゴールを踏まえ、最適な解決策を適用し、適切に作動しているか評価・見直し続けることが各主体に期待される。
- 各主体がAIガバナンスの構築において留意する観点としての行動目標及び、実践のポイント並びに実践例を述べる。
- 行動目標は、一般的かつ客観的な目標であり、・・・実践のポイント及び実践例の採否は、各主体に委ねられる。
- 採用する場合であっても、各主体の事情に応じた修正及び取捨選択の検討が期待される。

分類	行動目標
1.環境・リスク分析	1-1 便益/リスクの理解 1-2 AI の社会的な受容の理解 1-3 自社の AI 習熟度の理解
2.ゴール設定	2-1 AI ガバナンス・ゴールの設定
3.システムデザイン	3-1 ゴール及び乖離の評価及び乖離対応の必須化 3-2 AI マネジメントの人材のリテラシー向上 3-3 各主体間・部門間の協力による AI マネジメント強化 3-4 予防・早期対応による利用者のインシデント関連の負担軽減
4.運用	4-1 AI マネジメントシステム運用状況の説明可能な状態の確保 4-2 個々の AI システム運用状況の説明可能な状態の確保 4-3 AI ガバナンスの実践状況の積極的な開示の検討
5.評価	5-1 AI マネジメントシステムの機能の検証 5-2 社外ステークホルダーの意見の検討
6.環境・リスクの再分析	6-1 行動目標 1-1～1-3 の適時の再実施

https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_3.pdf

Ⅱ. セキュリティの確保

1. セキュリティ確保について

- **主体毎に重要となる事項**
- 別添本文4頁
- https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_3.pdf

	第 2 部. C.共通の指針	「共通の指針」に加えて主体毎に重要となる事項		
		第 3 部. AI 開発者 (D)	第 4 部. AI 提供者 (P)	第 5 部. AI 利用者 (U)
1) 人間中心	① 人間の尊厳及び個人の自律 ② AI による意思決定・感情の操作等への留意 ③ 偽情報等への対策 ④ 多様性・包摂性の確保 ⑤ 利用者支援 ⑥ 持続可能性の確保	-	-	-
2) 安全性	① 人間の生命・身体・財産、精神及び環境への配慮 ② 適正利用 ③ 適正学習	i．適切なデータの学習 ii．人間の生命・身体・財産、精神及び環境に配慮した開発 iii．適正利用に資する開発	i. 人間の生命・身体・財産、精神及び環境に配慮したリスク対策 ii. 適正利用に資する提供	i. 安全を考慮した適正利用
3) 公平性	① AI モデルの各構成技術に含まれるバイアスへの配慮 ② 人間の判断の介在	i. データに含まれるバイアスへの配慮 ii. AI モデルのアルゴリズム等に含まれるバイアスへの配慮	i. AI システム・サービスの構成及びデータに含まれるバイアスへの配慮	i. 入力データ又はプロンプトに含まれるバイアスへの配慮
4) プライバシー保護	① AI システム・サービス全般におけるプライバシーの保護	i．適切なデータの学習（D-2） i. 再掲）	i. プライバシー保護のための仕組み及び対策の導入 ii. プライバシー侵害への対策	i. 個人情報の不適切入力及びプライバシー侵害への対策
5) セキュリティ確保	① AI システム・サービスに影響するセキュリティ対策 ② 最新動向への留意	i. セキュリティ対策のための仕組みの導入 ii. 最新動向への留意	i. セキュリティ対策のための仕組みの導入 ii. 脆弱性への対応	i. セキュリティ対策の実施

2. セキュリティ確保に関する記載

- 本編の「共通の指針」は、前述のとおり10の指針からなりますが、そのうち、**セキュリティの確保**については、本編18頁において以下のとおり記載されています。

- 各主体は、AIシステム・サービスの開発・提供・利用において、不正操作によってAIの振る舞いに意図せぬ変更又は停止が生じることのないように、セキュリティを確保することが重要である。

- **①AIシステム・サービスに影響するセキュリティ対策**
  - ✧ AIシステム・サービスの機密性・完全性・可用性を維持し、常時、AIの安全安心な活用を確保するため、その時点での技術水準に照らして合理的な対策を講じる
  - ✧ AIシステム・サービスの特性を理解し、正常な稼働に必要なシステム間の接続が適切に行われているかを検討する
  - ✧ 推論対象データに微細な情報を混入させることで関連するステークホルダーの意図しない判断が行われる可能性を踏まえて、AIシステム・サービスの脆弱性を完全に排除することはできないことを認識する

- **② 最新動向への留意**
  - ✧ AIシステム・サービスに対する外部からの攻撃は日々新たな手法が生まれており、これらのリスクに対応するための留意事項を確認する

3. AIによる便益/リスク

- 別添では、「AIによる便益/リスク」について解説しているが、セキュリティに関連するとリスクとして、以下の2つが挙げられている。

- **データ汚染攻撃等のAIシステムへの攻撃**
  - AIの学習実施時では性能劣化及び誤分類につながるような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、AIの推論結果又はAIへの指示であるプロンプトを通した攻撃等もリスクとして存在する。例えば、とあるチャットボットでは、悪意のある集団による人種差別的な質問の組織的な学習により、ヘイトスピーチを繰り返し発言するようになった
  - 間接プロンプトインジェクションやマルウェアの生成など、悪意のある第三者によりRAGが悪用されるリスクがある（1.1版で加筆）

- **機密情報の流出**
  - AIの利用においては、個人情報及び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。例えば、AIサービスの利用上、従業員が業務利用のため、機密情報に該当するソースコードを、業務外利用者向けの対話型生成AIに入力してしまう事例が明らかになっている。生成AI系サービスでは利用障壁が下がっていることから、特に、企業のルール等が未整備の場合、企業による管理の外で、従業員が業務に業務外利用者向けの生成AIを用い、リスクの高い使用をする恐れもある。またRAGの活用など外部のサービス・データ等と連携する場合は、意図しない範囲に重要情報（個人情報・機密情報等）が漏洩してしまうこと等に特に注意が必要となる。加えて、生成AIを用いて生成したプログラムコードにセキュリティ上の脆弱性等が含まれていた場合、情報の改ざんや漏洩等につながる恐れがある。（1.1版で加筆）
  - ただし、エンタープライズグレードのセキュリティ機能が組み込まれた、ビジネス利用を想定した対話型生成AIも存在する。企業は、特に機密情報を処理するにあたっては、代わりにそのようなサービス又はアプリケーションを使用することが推奨される

- **こうしたリスクに対応すべく、「AI事業者ガイドライン」の対象者ごとに期待されるセキュリティ確保に関して、以下のとおり、具体的な方法が説明されています。**

4．AI開発者向け　セキュリティ確保のポイント

- AI開発者が講じておくべき、各開発フェーズにおけるセキュリティ確保のための対策としては、以下の説明がなされています。

- **<u>１　データ前処理・学習時</u>**
- **適切なデータの学習**
  - ✧ プライバシー・バイ・デザイン等を通じて、学習時のデータについて、適正に収集するとともに、第三者の個人情報、知的財産権に留意が必要なもの等が含まれている場合には、法令に従って適切に扱うことを、AIのライフサイクル全体を通じて確保する（安全性、プライバシー保護にも関連）
  - ✧ 学習前・学習全体を通じて、データのアクセスを管理するデータ管理・制限機能の導入検討を行う等、適切な保護措置を実施する
  - ➢ データ前処理・学習時についてあげられているのは、「AI事業者ガイドライン」の対象者のうちAI開発者のみに求められる対策です。
  - ➢ 別添では、適切なデータ学習のポイントとともに、**具体的な手法**として、以下が紹介されています。
    - データに個人情報、機密情報、著作権等の権利又は法律上保護される利益に関係するものが含まれていないか、確認を実施
    - 権利又は法律上保護される利益に関係するものが含まれる場合には、個人情報・機密情報・著作権等の適切な取扱いを実施
    - データが適切（正確性及び完全性等の品質が確保されている）かつ安全であることを保証するための対策を実施
    - 技術的に可能で合理的な範囲で、データの出所を把握するための手段の実施

4．AI開発者向け　セキュリティ確保のポイント　（続き）

- **２　AI開発時**
- **セキュリティ対策のための仕組みの導入**
  - ✧ AIシステムの開発の過程を通じて、採用する技術の特性に照らし適切にセキュリティ対策を講ずる（セキュリティ・バイ・デザイン）
  - ➢ 別添では、当該仕組みの導入のポイントとともに、**具体的な手法**として、①セキュリティ・バイ・デザインに基づくセキュリティ対策の実施、②セキュリティ対策の強化について、詳細な解説がされています。
- また、「機械学習利用システムの被害と脅威の例」という図表において、被害の内容ごとに、それを引き起こす脅威が掲載されており参考になります。
- 「**セキュリティ対策のための仕組みの導入**」については、AI提供者によるAIシステム実装時にも期待されています。

機械学習利用システムの被害と脅威の例

被害の内容			被害を引き起こす脅威	
			機械学習特有の脅威	その他の脅威
完全性又は可用性の侵害	システムの誤作動	意図に反する機械学習要素の動作による	データポイズニング攻撃	機械学習要素を実装するソフトウェア・ハードウェアに対する従来型の攻撃
			モデルポイズニング攻撃	
			汚染モデルの悪用	
			回避攻撃	
		その他の要因による		システムに対する従来型の攻撃
	計算資源の浪費	機械学習要素による	データポイズニング攻撃（資源枯渇型）	機械学習要素を実装するソフトウェア・ハードウェアに対する従来型の攻撃
			モデルポイズニング攻撃（資源枯渇型）	
			汚染モデルの悪用	
			スポンジ攻撃	
		その他の要因による		システムに対する従来型の攻撃
機密性の侵害	AIモデルについての情報の漏洩		モデル抽出攻撃	AIモデルを窃取する従来型の攻撃
	訓練用データに含まれるセンシティブ情報の漏洩		訓練用データに関する情報漏洩攻撃	データを窃取する従来型の攻撃
			データポイズニング攻撃（情報埋込型）	
	その他の機密情報の漏洩		モデルポイズニング攻撃（情報埋込型）	

出典
「AI事業者ガイドライン（第1.1版）別添（付属資料）」（令和7年3月28日）99頁

4．AI開発者向け　セキュリティ確保のポイント　（続き）

- **３　AI開発後**
- **最新動向への留意**
  - ✧ AIシステムに対する攻撃手法は日々新たなものが生まれており、これらのリスクに対応するため、開発の各工程で留意すべき点を確認する。
- **関連するステークホルダーへの情報提供**（主として「透明性」に関することではあるが・・・）
  - ✧ 自らの開発するAIシステムについて、例えば以下の事項を適時かつ適切に関連するステークホルダーに（AI提供者を通じて行う場合を含む）情報を提供する
    - （中略）
    - AI モデルで学習するデータの収集ポリシー、学習方法及び実施体制等（公平性、プライバシー保護にも関連）
- なお、**高度なAIシステムの開発者**には、遵守事項について、プラスアルファの記載がなされている点に留意が必要。
  - 具体的には、別添120頁において、「最先端の基盤モデル及び生成AIシステムを含む、高度なAIシステムを開発するAI開発者については、以下の「高度なAIシステムを開発する組織向けの広島プロセス国際行動規範」を遵守すべきである」と記載されています。
  - セキュリティの確保に関する事項も載っており、高度なAIシステムの開発者は一読が必須。
  - 詳細については、「高度なAIシステムを開発する組織向けの広島プロセス国際行動規範」参照。

5. AI提供者向け　セキュリティ確保のポイント

- 「AI事業者ガイドライン」では、AI提供者について、AIシステムをアプリケーション、製品、既存のシステム、ビジネスプロセス等に組み込んだサービスとしてAI利用者等に提供する事業者と定義しています。
- AI提供者が講じておくべき、各フェーズにおける「セキュリティ確保」のための対策としては、以下の説明がなされています。

- **1　AIシステム実装時**
  - **セキュリティ対策のための仕組みの導入**
  - AIシステム・サービスの提供の過程を通じて、採用する技術の特性に照らし適切にセキュリティ対策を講ずる（セキュリティ・バイ・デザイン）。
  - ➢AIのセキュリティに留意し、AIシステムの機密性・完全性・可用性を確保するため、その時点での技術水準に照らして合理的な対策を講ずることが期待される。＊RAGの活用
  - ➢セキュリティが侵害された場合に講ずるべき措置について、当該AIシステムの用途、特性、侵害の影響の大きさ等を踏まえ、あらかじめ整理しておくことが期待される。
  - 別添では、当該仕組みの導入のポイントとともに、**具体的な手法**として、以下が解説されています。
    - ① セキュリティ・バイ・デザインにもとづくセキュリティ対策の実施
    - ② AIに対する攻撃の分類（システムの誤作動、AIモデル情報の漏洩、訓練用データに含まれるセンシティブ情報の漏洩）
    - ③ セキュリティ侵害発生時の措置の検討

5.　AI提供者向け　セキュリティ確保のポイント　（続き）

- **2　AIシステム・サービス提供後**

  - P-5） ii .**脆弱性への対応**

✧ AIシステム・サービスに対する攻撃手法も数多く生まれているため、最新のリスク及びそれに対応するために提供の各工程で気を付けるべき点の動向を確認する。また、脆弱性に対応することを検討する。

  - 別添ではそのポイントとともに**具体的な手法**として、以下をあげて解説しています。
    - ①AIモデルに対する脆弱性に関するリスクへの留意
    - ②機械学習特有の各種攻撃への対策

  - **「関連するステークホルダーへの情報提供」**（主として「透明性」に関することではあるが・・・）

✧ 提供するAIシステム・サービスについて、例えば以下の事項を平易かつアクセスしやすい形で、適時かつ適切に情報を提供する。

    - （中略）
    - AIモデルにて学習するデータの収集ポリシー、学習方法、実施体制等（公平性、プライバシー保護にも関連）

P-6） ii．関連するステークホルダーへの情報提供

✧ **提供するAIシステム・サービスについて、例えば以下の事項を平易かつアクセスしやすい形で、適時かつ適切に情報を提供する（「6）透明性」）**

- （省略）
- AIモデルにて学習するデータの収集ポリシー、学習方法、実施体制等（「3）公平性」、「4）プライバシー保護」、「5）セキュリティ確保」）

[ポイント]

- AI提供者は、個人の権利・利益に重大な影響を及ぼす可能性のある分野において AI を利用する場合等、AIを活用する際の社会的文脈を踏まえ、AI利用者の納得感及び安心感の獲得、また、そのための AI の動作に対する証拠の提示等を目的として、AIの出力結果の説明可能性を確保することが期待される。その際、どのような説明が求められるかを分析・把握し、必要な対応を講じることが期待される。
- リスクを評価して対処した後、AI システムが規制、AIガバナンス及び倫理基準に準拠しているかどうかを検証し、関連するステークホルダーと共有することが重要となる。これにより、リスク又は意思決定及び行動の背後にある論理的根拠の理解が促進される。モニタリング及びレビューのプロセス並びにツールを確立するだけでなく、定期的なコミュニケーション及びレビューを実施して、AIの望ましくない動作及びインシデントに関する情報についても関連するステークホルダーと確実に共有することが重要である。

[具体的な手法]

- AIシステム・サービスについての情報共有
  - 提供するAIシステム・サービスがAIを用いたものであること及びその用途・方法、AIを活用している範囲、AIの性質、利用の態様等に応じた便益及びリスク
  - 提供するAIシステム・サービスの活用の範囲・方法に関する定期的な確認方法（特に、AIシステムが自律的に更新される場合の観測及び確認方法）、確認の重要性・頻度、未確認によるリスク等
  - 活用の過程における、AIの機能を向上させ、リスクを抑制するために実施するAIシステムのアップデート、点検、修理等
  - 安全性、セキュリティ、並びに社会的リスク及び人権に対するリスクについて実施された評価の詳細
  - 適切な使用領域、その使用に影響を及ぼすAIモデル又はAIシステム・サービスの能力及び性能上の限界
  - 有害な偏見、差別、プライバシー侵害の脅威、公平性への影響等、AIモデル又はAIシステム・サービスが安全性や社会に及ぼす影響及びリスクについての議論及び評価
  - 開発段階以降のAIモデル又はAIシステム・サービスの適合性を評価するために実施されたレッドチーミングの結果
  - 情報提供に際しての留意点（適切なタイミング、提供すべき情報を利用前に提供、それができない場合の体制整備）

6-1．AI利用者向け　セキュリティ確保のポイント

- AI利用者については、「セキュリティ確保」のための事項として、AIシステム・サービス利用時のセキュリティ対策の実施があげられており、具体的には、以下の行動が期待されています。

  - ①　AI提供者によるセキュリティ上の留意点を遵守する
  - ②　AIシステム・サービスに機密情報等を不適切に入力することがないよう注意を払う

- **【ポイント】**

- AI利用者は、**セキュリティが侵害された場合に講ずるべき措置**について、AI提供者から情報提供（AI開発者の情報も含む）があった場合には、AIシステム・サービスの利用にあたり留意することが望ましい。また、AIシステム・サービスを利用するにあたり、セキュリティ上の疑問を感じた場合は、AI提供者（又はAI提供者を通じてAI開発者）にその旨を報告することが期待される。
- AI利用者は、業務外利用者側でセキュリティ対策を実施することが想定されている場合には、AI提供者からの情報提供（AI開発者の情報も含む）を踏まえ、AIシステムのセキュリティに留意し、業務外利用者と連携して必要なセキュリティ対策を講ずることが期待される。

6-2.　AI利用者向け　セキュリティ確保の具体的な手法

- **脆弱性に関するリスクの認識**
  - 学習が不十分であること等の結果、AIモデルが正確に判断することができるデータに、人間には判別できない程度の微少な変動を加え、そのデータを入力すること等により、作為的にAIモデルが誤動作するリスク（例：Adversarial example攻撃）
  - 教師あり学習において不正確なラベリング等がなされたデータを混在させることで、誤った学習が行われるリスク
  - AIモデルが容易に複製されるリスク
  - AIモデルから学習に用いられたデータをリバースエンジニアリングされるリスク

- **セキュリティ侵害発生時の措置の検討**
  - 初動措置
    - AIシステムのロールバック、代替システムの利用等による復旧
    - AIシステムの停止（キルスイッチ）
    - AIシステムのネットワークからの遮断
    - セキュリティ侵害の内容の確認
    - 関連するステークホルダーへの報告
  - 補償・賠償等を円滑に行うための保険の利用
  - 第三者機関の設置及びその機関による原因調査・分析・提言

6-2. AI利用者向け　セキュリティ確保の具体的な手法　（続き）

- **機密情報等を含むプロンプトの入力**
  - 例えば、生成 AI サービスの利用にあたって、入力する機密情報が生成 AI サービスの提供者においてAIの学習データとして利用されることが予定されている場合には、機密情報を含むプロンプトを入力しないよう留意する
  - AIに入力する情報への留意
    - AIに過度に感情移入すること等により、機密情報をむやみにAIに与えないようにする

Ⅲ.AIのセキュリティ確保のための技術的対策に係るガイドライン（案）

本ガイドラインの策定の背景

AI の安全・安心な活用促進に関しては、「AI 事業者ガイドライン」（総務省・経済産業省）が策定され、各主体が連携して取り組むべき共通の指針の一つとして「セキュリティ確保」が位置付けられている。また、AI の安全性に対する国際的な関心の高まりを踏まえ、令和５年の日本議長国下の G7 において生成 AI 等に関する国際ルールの検討を行う「広島 AI プロセス」が立ち上げられ、安全・安心で信頼できる AI を実現するためのルール作りを日本が主導しているほか、「統合イノベーション戦略 2024」（令和６年６月４日閣議決定）に基づき、我が国においても関係省庁・関係機関から構成される「AI セーフティ・インスティテュート（AISI)」が設立され、AI に対する脅威の特定や、レッドチーミングガイドの策定等が行われてきている。

「デジタル社会の実現に向けた重点計画」（令和７年６月 13 日閣議決定）では、総務省が、令和７年度末までに、生成 AI とセキュリティのガイドラインを策定・公表することとされているほか、「サイバーセキュリティ 2025」（令和７年６月 27 日サイバーセキュリティ戦略本部決定）においても、AI の安心・安全な開発・提供に向けたセキュリティガイドラインを策定することとされている。

総務省では、このような状況を踏まえ、令和７年９月から「サイバーセキュリティタスクフォース」の下に「AI セキュリティ分科会」を開催し、同年 12 月に取りまとめをいただいたところである。本ガイドラインは、当該取りまとめの内容を踏まえ、AI のセキュリティ確保のための技術的対策例を示すものとして策定するものである。

本ガイドラインの内容は、策定時点の状況が反映されているに過ぎず、AI の技術進展が著しい中にあって、AI 開発者及び AI 提供者においては、新たな脅威や技術の進展に応じた対応を不断に検討していくことが重要である。

ガイドライン案　目次

本ガイドラインの位置付け

本ガイドラインは、AI 事業者ガイドラインで示された共通の指針、「AI セーフティに関する評価観点ガイド」（AISI）で示された「AI セーフティにおける重要要素」及び「AI セーフティ評価の観点」を踏まえ、AI の「セキュリティ確保」を取り扱う[3]。

本ガイドラインにおいては、AI の「セキュリティ確保」として、「不正操作による機密情報の漏えい、AI システムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理している。

[3] なお、「AI セーフティに関する評価観点ガイド」（AISI）では、「AI セーフティ」及び「セキュリティ確保」について以下のとおり記載されている。

AI セーフティ
人間中心の考え方をもとに、AI 活用に伴う社会的リスク※を低減させるための安全性・公平性、個人情報の不適正な利用等を防止するためのプライバシー保護、AI システムの脆弱性等や外部からの攻撃等のリスクに対応するためのセキュリティ確保、システムの検証可能性を確保し適切な情報提供を行うための透明性が保たれた状態。　（出典：総務省・経済産業省「AI 事業者ガイドライン（第 1.0 版）」）
※社会的リスクには、物理的、心理的、経済的リスクも含む（出典：Department for Science, Innovation and Technology, UK AISI “Introducing the AI Safety Institute”）

3.6　セキュリティ確保
■評価観点の概要説明
LLM システムに対する悪意ある攻撃やヒューマンエラーによる設定ミス等の影響を最小限にとどめるために、セキュリティ確保は重要である。（中略）LLM システム全体の脆弱性に対策し、不正操作による機密情報の漏えい、LLM システムの意図せぬ変更または停止が生じないような状態を目指す。

対象とする主な脅威

本ガイドラインでは、攻撃の具体的な可能性が比較的高いと考えられるプロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）への対策を主に示す。これらの攻撃は基本的にプロンプトの入力により実施可能であるため、攻撃の具体的な可能性が比較的高いと考えられる。

なお、一般的には対策を講じるべき脅威の特定には、以下の要素を考慮して、個別の事例ごとに検討することになると考えられる[5]。

1）脅威の影響の大きさ
アプリケーションの用途によって、インシデント発生時の影響の性質（例えば、事業停止による損失、信用の失墜など）、範囲、深刻さの度合いは異なり、それ故にリスクの大きさも異なるため、対策の優先度は異なる。

2）脅威が発生する可能性
攻撃者が攻撃を実行できる可能性や、AIシステムがおかれた環境においてインシデントが起こり易いか否かで、対策の優先度は異なる。

脅威の内容

2.1.1　プロンプトインジェクション攻撃[6]

プロンプトインジェクション攻撃とは、LLM に細工をした入力を行うことで、不正な出力をさせる攻撃である。本ガイドラインにおいて、LLM に細工をしたプロンプトを入力することで実施するものを直接プロンプトインジェクション攻撃といい、LLM に細工をしたデータを参照させることで実施するものを間接プロンプトインジェクション攻撃という。

「不正な出力」の例としては、以下が挙げることができる。

- 本来は開示すべきではない、RAG 用のデータストア（ベクトルデータベースやファイルシステム等）の内容を含む出力をさせる
- 連携するシステムを不正操作するコード（SQL クエリやシステムコマンド等）を LLM に生成させ、これを連携するシステム上で実行させることで、データベースやシステムからの機密情報の漏えいや、データの改ざん・削除等を行う
- 本来は開示すべきではない、LLM の内部設定が記載されたシステムプロンプトを含む出力をさせる
- ユーザが LLM を利用する目的が果たされなくなるような誤った内容を出力させる

2.1.2　DoS 攻撃（サービス拒否攻撃）

DoS 攻撃（サービス拒否攻撃）とは、LLM に、AI システムが膨大な処理を必要とするプロンプト入力を行うことで、AI システムへの想定以上の計算負荷や、経済的な損失を生じさせ、AI システムの応答の遅延・停止を引き起したり、サービスの継続性を損なわせたりする攻撃である[7]。

その他の脅威

2.1で掲げたプロンプトインジェクション攻撃やDoS攻撃（サービス拒否攻撃）はプロンプト入力のみを介して実行することも可能である。このほか、単純なプロンプト入力ではなく、予めデータを汚染させるなど攻撃に一定の前提条件が必要となるものや、攻撃に当たってLLMへの執拗なアクセスが必要となるものとして、以下の脅威もある。

- **データポイズニング攻撃**

  データポイズニング攻撃とは、基盤モデルやLLMが学習するデータに細工をし、LLMに不正な出力をさせる攻撃である。攻撃者は、細工をしたデータを用意し、これを何らかの手段によって、事前学習データやファインチューニングデータに入れ込むことで、LLMが特定のプロンプト入力に対して不正な回答を出力するようにしてしまう。

- **細工をしたモデルの導入を通じた攻撃**

  細工をしたモデルの導入を通じた攻撃とは、細工をしたLLMをAIシステムに組み込ませ、LLMに不正な動作をさせる攻撃である。攻撃者は、細工をしたLLMを用意し、これを外部に提供することで、細工をしたLLMをAIシステムに組み込ませ、AIシステムが不正な動作をするようにしてしまう。

- **モデル抽出攻撃**

  モデル抽出攻撃とは、LLMに繰り返しアクセスし、LLMが出力する各単語とその出現確率を分析することで、当該LLMと類似のLLMを複製する攻撃である。これにより、当該LLMに係る競争上の地位低下や、当該LLMに含まれる機密情報の窃取などにつながる。

Ⅳ.システム開発におけるセキュリティ関連裁判例

①東京地判　平成26年1月23日　一部認容　平成23年（ワ）32060号

- 第二　事案の概要
- 　本件は、原告が、被告との間で、原告のウェブサイトにおける商品の受注システムの設計、保守等の委託契約を締結したところ、被告が製作したアプリケーションが脆弱であったことにより上記ウェブサイトで商品の注文をした顧客のクレジットカード情報が流失し、原告による顧客対応等が必要となったために損害を被ったと主張して、被告に対し、上記委託契約の債務不履行に基づき損害賠償金一億〇九一三万五五二八円及びこれに対する訴状送達の日の翌日である平成二三年一〇月一五日から支払済みまで商事法定利率年六分の割合による遅延損害金の支払を求める事案である。

契約の内容

- 三　争点〈２〉（被告の債務不履行責任の有無）について
- （１）　原告と被告との間の契約関係
- 被告が負うべき債務の内容を判断する前提として、原告と被告との間の契約関係について検討すると、前提事実及び前記認定の事実によれば、被告は、原告との間で、本件基本契約を締結した上で、個別契約として、本件システムの製作（本件システム発注契約）、保守サービス（一年ごとに更新）、クレジットカード情報の把握（金種指定詳細化）、本件ウェブサイトのデザイン変更作業（本件ウェブサイトメンテナンス契約）等に係る本件個別契約を締結したのであるから、個別契約ごとに、当該個別契約及び本件基本契約に基づく債務を負うものと認められる（本件基本契約二条により、個別契約には本件基本契約が適用される。）。
- これに対し、原告は、被告との間で締結した本件基本契約（同日に締結した覚書を含む。）、本件ウェブサイトメンテナンス契約及び本件基本契約に基づく各個別契約は全て一体の契約としてみるべきであると主張するが、本件基本契約及び本件個別契約は別の時期に締結されたものであり、個別契約ごとに内容も異なるのであるから、これらの契約を全て一体の契約としてみて、本件個別契約に基づき発生する債務を一体として把握することはできないから、原告の上記主張は採用できない。
- （２）　そして、前記二のとおり、本件流出の原因はＳＱＬインジェクションであると認められるから、本件個別契約及び本件基本契約に基づき、被告に債務不履行一、三及び五が認められるかを検討する。
- ア　債務不履行一（適切なセキュリティ対策が採られたアプリケーションを提供すべき債務の不履行）
- （ア）　前提事実のとおり、被告は、平成一一年一月四日に本件システム発注契約を締結して本件システムの発注を受けたのであるから、その当時の技術水準に沿ったセキュリティ対策を施したプログラムを提供することが黙示的に合意されていたと認められる。そして、本件システムでは、金種指定詳細化以前にも、顧客の個人情報を本件データベースに保存する設定となっていたことからすれば、被告は、当該個人情報の漏洩を防ぐために必要なセキュリティ対策を施したプログラムを提供すべき債務を負っていたと解すべきである。

義務の内容

- （２）　そして、前記二のとおり、本件流出の原因はＳＱＬインジェクションであると認められるから、本件個別契約及び本件基本契約に基づき、被告に債務不履行一、三及び五が認められるかを検討する。
- ア　債務不履行一（適切なセキュリティ対策が採られたアプリケーションを提供すべき債務の不履行）
- （ア）　前提事実のとおり、被告は、平成二一年二月四日に本件システム発注契約を締結して本件システムの発注を受けたのであるから、その当時の技術水準に沿ったセキュリティ対策を施したプログラムを提供することが黙示的に合意されていたと認められる。そして、本件システムでは、金種指定詳細化以前にも、顧客の個人情報を本件データベースに保存する設定となっていたことからすれば、被告は、当該個人情報の漏洩を防ぐために必要なセキュリティ対策を施したプログラムを提供すべき債務を負っていたと解すべきである。
- そこで検討するに、«証拠略»によれば、経済産業省は、平成一八年二月二〇日、「個人情報保護法に基づく個人データの安全管理措置の徹底に係る注意喚起」と題する文書において、ＳＱＬインジェクション攻撃によってデータベース内の大量の個人データが流出する事案が相次いで発生していることから、独立行政法人情報処理推進機構（以下「ＩＰＡ」という。）が紹介するＳＱＬインジェクション対策の措置を重点的に実施することを求める旨の注意喚起をしていたこと、ＩＰＡは、平成一九年四月、「大企業・中堅企業の情報システムのセキュリティ対策～脅威と対策」と題する文書において、ウェブアプリケーションに対する代表的な攻撃手法としてＳＱＬインジェクション攻撃を挙げ、ＳＱＬ文の組み立てにバインド機構を使用し、又はＳＱＬ文を構成する全ての変数に対しエスケープ処理を行うこと等により、ＳＱＬインジェクション対策をすることが必要である旨を明示していたことが認められ、これらの事実に照らすと、被告は、平成二一年二月四日の本件システム発注契約締結時点において、本件データベースから顧客の個人情報が漏洩することを防止するために、ＳＱＬインジェクション対策として、バインド機構の使用又はエスケープ処理を施したプログラムを提供すべき債務を負っていたということができる。
- そうすると、本件ウェブアプリケーションにおいて、バインド機構の使用及びエスケープ処理のいずれも行われていなかった部分があることは前記二のとおりであるから、被告は上記債務を履行しなかったのであり、債務不履行一の責任を負うと認められる。

②東京地判　令和2年10月13日　請求棄却　平成28年（ワ）10775号

- 第２　事案の概要
- 　　本件は、原告が、被告との間で締結された、原告が運営する電子商取引を行うウェブサイト（以下「ＥＣサイト」という。）の製作に係る請負契約及び同ＥＣサイトの保守管理契約に関し、被告において、〈１〉上記請負契約に基づく、同ＥＣサイトの顧客のクレジットカード情報を保持しない仕様でソフトウェアを製作する義務、〈２〉上記請負契約に基づく、その締結当時の技術水準に沿ったセキュリティ対策を施したソフトウェアを提供する義務、〈３〉上記保守管理契約に基づく、上記〈１〉及び〈２〉と同様の義務を負っていたところ、被告が当該〈１〉～〈３〉の義務の少なくとも一つを怠ったことから、同ＥＣサイトに第三者が侵入して、原告の顧客のクレジットカード情報が漏洩する結果となった旨を主張して、被告に対し、上記請負契約に係る瑕疵の修補に代わる損害賠償請求権及び同契約の債務不履行による損害賠償請求権又は上記保守管理契約の債務不履行による損害賠償請求権に基づき、原告の損害相当額である７８４３万６０１７円及びこれに対する本件訴状の送達日の翌日である平成２８年４月１２日から支払済みまで商事法定利率（平成２９年法律第４５号による改正前の商法５１４条に基づくもの）年６分の割合による遅延損害金を支払うことを求める事案である。

争点２の判断

- ３　**争点２（被告が本件サイトの顧客のクレジットカード情報を保持しない仕様で本件ソフトウェアを製作する義務に違反したか否か**）について
- 　　(1)　前記前提事実のとおり、本件決済モジュールは、Ｈ社が開発し、Ｇ社が提供していたものであるから、被告は、本件決済モジュールの設計を行う開発者には該当しないというべきである。
- 　　(2)　前記認定事実に照らし、本件請負契約に基づく被告の具体的な業務内容は、本件請負注文書の内容によって定まっていたものと認められる。本件請負注文書において、被告は、本件サイトにおけるクレジットカード決済機能を「導入」するものとされ、同機能を開発し、又は同機能を提供するプログラムを製作するものとはされていなかった。そして、前記認定事実のとおり、原告は、被告に対し、本件サイトのショッピングカート機能について、旧カートシステムをＢ独自のシステムへと変更したいとの意向を伝えた上、４万円（消費税別）の代金に相当する作業を請け負わせたもので、本件請負注文書の他の項目をみても、被告が本件サイトにクレジットカード決済機能を提供する何らかのプログラムを開発する旨の記載は存在しない。そうすると、被告は、本件決済モジュールの開発を行う開発者には該当しないというべきである。

続き

- (4)　前記前提事実のとおり、本件決済モジュールは、ＰＨＰ言語で記述され、ＧＰＬライセンスに従ってソースコードが公開されていたプログラムであるが、Ｂを利用したＥＣサイトにおいて、本件決済モジュールを用いてクレジットカード決済を行った場合、本件決済ログにクレジットカード情報が保存される実装になっていたことは、本件情報漏洩の判明後に発覚したもので、被告が原告に本件サイトを引き渡すまでの間に、当該不具合の存在が公になっていたものではない。また、ソフトウェアのソースコードが公開されていることや、ソフトウェアの不具合の具体的な内容が特定され、その原因が判明した時点において当該ソースコードの修正が容易であることは、当該不具合の調査が容易であることを何ら意味するものではない。むしろ、ソフトウェアのどの部分にいかなる不具合やセキュリティ脆弱性が潜んでいるかを調査するためには、高度の専門的知見と相当のコストを要することが通常であり、現に、原告は、本件決済モジュールを含む本件ソフトウェアの調査を、専門業者であるＩ社に依頼し、高額の調査費用を支払っている。
- (4)　以上の事情を総合すれば、被告が、本件決済モジュールの設計、開発及びカスタマイズを行う開発者に該当しないにもかかわらず、相当額の対価の支払を受ける約定もないのに、高度の専門的知見と相当のコストを要する作業を進んで請け負うことは考え難い。本件サイトに顧客のクレジットカード情報を保存しないことが、原告及び被告の共通認識となっていたとみられることを考慮しても、本件請負契約に関し、原告と被告との間で、被告が、本件決済モジュールのソースコードや、同モジュールが生成するログを調査し、同モジュールが、セキュリティ脆弱性を有しないか、異常処理を生じさせないかといった点を確認する義務を負うとの合意をしていたことを認めることはできない。この判断に反する原告の主張は、いずれも採用することができない。

争点3の判断

- **４　争点３（被告が本件請負契約の締結当時の技術水準に沿ったセキュリティ対策を施したソフトウェアを提供する義務に違反したか否か）について**
- （1）　本件基本契約に基づく被告の義務は、本件請負契約に基づく被告の義務ともなるものであるが、前記認定事実によれば、本件基本契約１１条は、被告において、請負業務の実施に当たって取り扱うことを認識している、原告管理下の顧客の個人情報を適切に取り扱わせるため、被告に善管注意義務を課した規定と解される。本件決済ログに暗号化された状態で保存されていた、本件サイトの顧客のクレジットカード情報は、被告において、これを取り扱うことを認識していなかった原告顧客の個人情報であるから、本件基本契約１１条の対象に含まれるものとはいえない。そうすると、本件基本契約１１条を根拠に、被告が本件決済モジュールに関し何らかの義務を負うということはできない。
- （2）　原告は、被告が、本件請負契約に基づき、瑕疵のない本件ソフトウェアを製作する義務の一環として、同契約の締結当時の技術水準に沿ったセキュリティ対策を施したソフトウェアを提供する義務を負っており、当該義務には、本件決済モジュールについて、本件仮想サーバー内にクレジットカード情報を保存しない仕様を実装させ、仮にこれと異なる実装になっている不具合があった場合には、そのような不具合を取り除いて、本件仮想サーバー内にクレジットカード情報を保存しないものへと修正する義務が含まれていた旨を主張する。
- しかし、前記２で認定説示したとおり、被告は、本件決済モジュールの設計、開発及びカスタマイズを行う開発者に該当せず、本件請負契約上、本件サイトにクレジットカード決済機能を導入することがその業務となっていたもので、同機能を開発し、又は同機能を提供するプログラムを製作することがその業務となっていたものではない。また、被告に勧められた経緯があったとはいえ、原告は、Ｇ社との間で決済代行サービスに係る加盟店契約を締結したのであるから、本件サイトにクレジットカード情報を保持する場合に、クレジットカード決済を取り扱う加盟店として、クレジットカード業界のセキュリティ基準であるＰＣＩ　ＤＳＳ（Payment Card Industry Data Security Standard。乙１９の２、乙２７、３２、３４。）に準拠すべき立場にあったのは、被告ではなく原告である。加盟店が、ＥＣサイトの構築を第三者に請け負わせるに当たり、自らが準拠すべきセキュリティ基準に沿ったものとすることについては、注文主の存在する加盟店から請負人に指示すべきものであり、注文主がそのような指示をしていないのに、請負人が黙示の合意の存在を理由に上記基準に沿ったＥＣサイトを構築する義務を負わされるとすれば、請負人の合理的な期待に反するといえない。さらに、前記認定事実のとおり、原告が締結した加盟店契約において、加盟店が、Ｇ社の決済代行サービスを利用するに当たり、同サービスに関連して同社が提供したコンピュータソフトウェアのプログラム等を改造又は変更する行為を禁止する旨の規定があることに照らせば、原告が、Ｇ社に無断で、本件決済モジュールの決済機能に関するソースコードを改造し又は変更することはできなかったものというべきである。本件において、原告は、本件請負契約に関し、本件決済モジュールの決済機能に関するソースコードを被告に改造させ又は変更させることにつき、Ｇ社から許諾を得ていなかったのであるから、この点をみても、被告が、これを改造し又は変更する義務を負っていたとは解されない。

③前橋地判　令和5年2月17日　一部認容　令和2年（ワ）145号/331号

- 第２　事案の概要
- １　原告は、被告との間で、平成２７年５月２１日、Ｘ市情報教育ネットワーク（Ｘ　Ｅｄｕｃａｔｉｏｎ　ＮＥＴｗｏｒｋ。以下「Ｘ・ＥＮＥＴ」という。）のデータセンターの移管設計及び構築業務に係る委託契約（以下「本件委託契約」という。）を締結し、その後、同年１０月１日、同データセンターの移管保守業務に係る委託契約（以下「本件保守契約」という。）を締結した。その後、何者かが、Ｘ・ＥＮＥＴの教育資料公開サーバにバックドア（ソフトウェアやシステムの一部として利用者に気付かれないよう秘密裏に仕込まれた正規の利用者認証やセキュリティ対策などを回避してこっそり遠隔操作するための窓口）を設置し、更に、何者かが、平成２９年１２月中旬頃から、上記のバックドアを利用して、上記のサーバ経由で内部ネットワークに侵入するようになり、不正ツールを保存し、更に、何者かが、平成３０年３月６日、多数の個人情報が保存されたファイル共有サーバから、各種ファイルを圧縮して教育資料公開サーバに収集、保存するなどしたこと（以下「本件不正アクセス」という。）により、多数の個人情報が流出した可能性が高いことが判明した。
- 本訴事件は、原告が、本件委託契約の受託者である被告には、Ｘ・ＥＮＥＴのデータセンターの移管設計及び構築業務を行うに当たり、ファイアウォール（管理者等が定めたルールに基づいて、ネットワークの間の通信を「許可」したり「拒否」したりする機能を有し、外部からの不正アクセスや攻撃を感知し、それらの記録を保存するセキュリティ機器）を適切に設定することにより通信制限を行う債務ないし注意義務があったにもかかわらず、被告がこれを怠り、これにより本件不正アクセスを許す事態となり、これに対応するために原告に多額の損害が生じたなどと主張して、選択的に、被告に対し、〈１〉債務不履行による損害賠償請求権に基づき、損害賠償金１億７７３５万６４４０円（ただし、請求原因として主張している損害の金額の合計は１億７７３３万７９９０円である。）及びこれに対する損害賠償の履行の請求をした日の翌日である平成３１年１月２９日から支払済みまで平成２９年法律第４５号による改正前の商事法定利率年６分の割合による遅延損害金の支払、又は、〈２〉不法行為による損害賠償請求権に基づき、損害賠償金１億７７３５万６４４０円（同上）及びこれに対する上記の業務に係るシステムを納品した日である平成２７年９月３０日から支払済みまで平成２９年法律第４４号による改正前の民法所定の年５分の割合による遅延損害金の支払を求めた事案である。
- 反訴事件は、被告が、原告は被告に対してモバイルルータの手配、Ｘ・ＥＮＥＴサーバのデジタルフォレンジック及び各教育機関に設置されたネットワーク対応ＨＤのクレンジング作業という各事務を委託し、被告がその各事務を処理するに当たり、モバイルルータ通信料５６３万４０７３円、Ｘ・ＥＮＥＴサーバ解析費用３８８万８００円及びＮＡＳクレンジング作業費用２０７万１１１１円の各費用が生じたと主張して、原告に対し、〈１〉主位的に、準委任契約による費用等の償還請求権に基づき、上記の各費用の合計１１５９万３１８４円及びこれに対する当該費用を請求した日の翌日である平成３１年４月１０日から支払済みまで平成２９年法律第４５号による改正前の商事法定利率年６分の割合による遅延損害金の支払を求め、〈２〉予備的に、商法５１２条の規定による商人の相当報酬請求権に基づき、報酬１１５９万３１８４円及びこれに対する上記と同様の遅延損害金の支払を求めた事案である。

争点

- **（１）　本訴事件**
- **ア　本件委託契約に基づく外部ファイアウォール及び内部ファイアウォールを適切に設定することにより通信制限を行うべき債務の不履行又は注意義務違反（不法行為）の有無**
- **イ　本件保守契約に基づく外部ファイアウォール及び内部ファイアウォールの設定における通信制限の不備を修正すべき債務の不履行又は注意義務違反（不法行為）の有無**
- **ウ　想定されるリスク及びその対策について適切な提案をすべき注意義務違反（不法行為）の有無**
- エ　債務不履行又は不法行為と本件不正アクセスとの間の相当因果関係の有無
- オ　原告の損害
- カ　被告の債務不履行に係る帰責事由の不存在（抗弁）
- キ　責任限定契約の適用の有無（抗弁）
- **（２）　反訴事件**
- ア　主位的請求（準委任契約に基づく費用等の償還請求）
- （ア）　原告と被告との間の準委任契約の成否
- （イ）　被告の準委任に係る事務を処理するのに必要な費用の額
- イ　予備的請求（商法５１２条の規定に基づく報酬請求権）
- （ア）　営業の範囲内における他人のための行為の有無及び当該行為の相当費用の額
- （イ）　無償で行うことの合意の成否（抗弁）

争点アの判断

- そこで検討するに、上記１（１）で認定したところからも明らかなとおり、ソフトウェアの開発に係る業務委託契約においては、契約締結の前後に提案依頼書、提案書、要件定義書、基本設計書などをやり取りすることにより委託業務の内容を確定していくものであるから、本件委託契約において受託者である被告が負う債務の内容は、同契約の契約書の記載の内容のみならず、同契約の前後にやり取りがされた要件定義書や基本設計書などの内容を総合的に考慮して確定すべきであると解するのが相当であるところ、上記１（１）ア（イ）、（ウ）、イ（ウ）、ウ（イ）、エ及びオで認定したところによれば、本件システムについて、〈１〉提案依頼書、提案書、要件定義書及び基本設計書には、外部ファイアウォール及び内部ファイアウォールにより通信制限を行うものと記載されていること、〈２〉設計方針及び基本設計書には、データセンター内理論接続図及び通信制限イメージにおいて、ＤＭＺネットワークと個人情報保護ネットワークとをつなぐ通信経路が存在しないことがそれぞれ認められ、これらの事実に加えて、〈３〉被告は、経験豊富な専門家を多数擁する技術的セキュリティ対策チームによる総合的なセキュリティソリューションを提供することを可能とする技術力を有していたことを併せ考えると、被告は、本件委託契約において、ＤＭＺネットワークと個人情報保護ネットワークとの間の通信経路を遮断するため、本件システムの提供に当たり、その外部ファイアウォール及び内部ファイアウォールを適切に設定して通信制限を行う債務を負っていたものと認めるのが相当である。

続き

- 　これに対し、被告は、原告が主張するような債務が契約書に明示されているわけではなく、適切に設定することによる通信制限を行う義務のような観念的抽象的な債務を負うものではない旨の主張をするが、上記に照らし、同主張は採用することができない。
- 　　イ　そして、〈１〉上記１（１）カのとおり、被告は、原告に対し、本件システムの外部ファイアウォールをＤＭＺネットワークから個人情報保護ネットワークを含む全ての内部ネットワークへの全ての通信を許可するとの設定及び内部ファイアウォールをＤＭＺネットワークを含む全てのネットワークからの個人情報保護ネットワークへの全ての通信を許可するとの設定としたまま、同システムを引き渡していること、〈２〉上記１（５）ウによれば、被告において、このような設定が不適切であったこと自体は自認していたことがそれぞれ認められ、これらの事実に加えて、本件記録を見ても、被告において、原告に対して外部ファイアウォール及び内部ファイアウォールを上記〈１〉のように設定したまま本件システムを引き渡した理由について合理的な理由がある旨の主張は見当たらないことを併せ考えると、被告には、本件システムの外部ファイアウォール及び内部ファイアウォールを適切に設定して通信制限を行う債務の不履行があるものと認めるのが相当である。

争点イの判断

- 原告は、被告には、本件保守契約に基づき、保守契約での対応を超えるものであった場合を除き、本件システムの不備を修正すべき債務又は注意義務があったところ、その債務の不履行又は注意義務の違反がある旨の主張をする。
- 　　しかし、〈１〉原告において、被告が、本件不正アクセスが発覚する以前において、原告市教委から通信制限の不備を指摘され、修正の要求を受けていたにもかかわらず、不備を自ら確認することもなく放置したと主張する点については、本件保守契約や保守手引書（甲５の１）の内容を見ても、その修正が本件保守契約の対象となっているものか否かについては判然としないといわざるを得ないから、被告に債務不履行又は注意義務違反があると直ちには認められないし、また、〈２〉原告において、被告が、他県での教育情報システムへの不正アクセスによる個人情報漏えい事件に関連して、原告市教委から問い合わせを受けたにもかかわらず、システムの確認を怠ったと主張する点については、本件記録を見ても、原告市教委が被告に対して具体的にどのような連絡をしたのかは判然としないから、被告に債務不履行又は注意義務違反があると直ちには認められないし、さらに、〈３〉原告において、被告が、その後も、原告から上記〈２〉とは別の通信制限の不備を指摘されたにもかかわらず、必要な措置を講じなかったと主張する点についても、同様に、本件記録を見ても、原告が被告に対して具体的にどのような連絡をしたのかは判然としないから、被告に債務不履行又は注意義務違反があると直ちには認められず、結局、原告の上記（１）の主張は採用することができない。

争点ウの判断

- 　原告は、被告を含む数社に対し、提案依頼書で「想定されるリスク及びその対策について適切な提案を行うこと」が可能な事業者を募集し、これに対し、被告が、提案書において、ＩＳＯ２７００１及びＩＳＯ９００１の認証取得者であることを強調し、原告の要求を十分に満たすことをアピールしたのであるから、被告は、想定されるリスク及びその対策について適切な提案をすべき注意義務がある旨の主張をする。
- 　しかし、上記１（１）アで認定したところによれば、〈１〉原告は、被告を含む数社に対し、提案依頼書により、想定されるリスク及びその対策について適切な提案をすることを求めていること、〈２〉被告は、原告に対し、提案書により、被告が強固なセキュリティ対策を講じることができることやＩＳＯ２７００１やＩＳＯ９００１などを取得していることをあげて、安全・安心面をアピールしていることがそれぞれ認められるが、これらの事実から直ちに、被告には、想定されるリスク及びその対策について適切な提案をすることに係る具体的な注意義務があったと認めることは困難というべきであるから、原告の上記（１）の主張は採用することができず、被告につき不法行為の成立は認められない。

Ⅴ.生成AIサービスにおけるセキュリティ対策と法的責任

どのような法的責任が考えられるか？

- 「AI事業者ガイドライン」 における「セキュリティの確保」に関する記載を中心に紹介しましたが、セキュリティの観点からは、共通の指針のうち「プライバシー保護」や「教育・リテラシー」の項目も重要。
- 「AI事業者ガイドライン」は、あくまで行政が示す将来の指針。法的拘束力はない。
  - 「AI事業者ガイドライン」自身、別添3頁において、「事業者の事業運営形態も様々であることが想定されることから、この付属資料を全て記載とおりに実施することが求められているものではない」と述べている。
- それでは、遵守しなくても良いか？
  - このような行政や業界団体のガイドラインは、司法判断の際も参考にされ得る。
- 具体的にどのような場面で、**法的責任**が問題になるか？
  - **契約上の義務違反または不法行為（過失）責任・・・過失の場合、権利侵害という結果に対する予見可能性、回避可能性が前提。**
  - 仮に、セキュリティの確保がされていないまたは不十分という理由で何らかの損害が発生した場合、AIの開発者、提供者、利用者は、その損害の被害者から、債務不履行または不法行為に基づく損害賠償請求等の訴えを提起される可能性があります。
  - ステークホルダーである企業の規模、人員のリソース等は一様ではなく、AIサービスの種類・内容、利用者の利用態様、加害者による具体的な侵害態様は事案によって異なり、権利侵害という結果の予見可能性、回避可能性も一様ではなく、その事案における具体的な事情に鑑みて判断されるべき。
  - **これまでのシステム関連裁判例を踏まえて、（生成）AIサービスにおけるセキュリティ対策と法的責任はどう考えるべきか？**

参考　著作権法30条の4

著作権法上　留意点

【学習段階又はプロンプト段階での著作物の入力行為】

（著作物に表現された思想又は感情の享受を目的としない利用）

- **第30条の4**
- 著作物は、次に掲げる場合**その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合**には、
- **その必要と認められる限度**において、いずれの方法によるかを問わず、利用することができる。
- ただし、当該著作物の種類及び用途並びに当該利用の態様に照らし**著作権者の利益を不当に害することとなる場合は、この限りでない。**

一　著作物の録音、録画その他の利用に係る技術の開発又は実用化のための試験の用に供する場合

二　情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう。第47条の5第1項第2号において同じ。）の用に供する場合

三　前２号に掲げる場合のほか、著作物の表現についての人の知覚による認識を伴うことなく当該著作物を電子計算機による情報処理の過程における利用その他の利用（プログラムの著作物にあっては、当該著作物の電子計算機における実行を除く。）に供する場合

「AI と著作権に関する考え方について」（令和6年3月15日　文化庁）37～38頁

- https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

- **ク 生成指示のための生成AIへの著作物の入力について**
  - 生成AIに対して生成の指示をする際は、プロンプトと呼ばれる複数の単語又は文章や、画像等を生成AIに入力する場合があり、入力に当たっては、著作物の複製等が生じる場合がある。
  - この生成AIに対する入力は、生成物の生成のため、入力されたプロンプトを情報解析するものであるため、これに伴う著作物の複製等については、法第30条の４の適用が考えられる。
  - ただし、**生成AIに対する入力に用いた既存の著作物と類似する生成物を生成させる目的で当該著作物を入力する行為**は、生成AIによる情報解析に用いる目的の他、入力した著作物に表現された思想又は感情を享受する目的も併存すると考えられるため、法第30条の４は適用されないと考えられる。

ご清聴ありがとうございました。

＜質問等はこちらまで＞

TMI総合法律事務所

〒106-6123　東京都港区六本木6丁目10番1号
六本木ヒルズ森タワー23階

柴野相雄　tshibano@tmi.gr.jp