SUMITOMO ELECTRIC
Connect with Innovation

CSO25-0048

AIセーフティ・AIマネジメントを取り巻く国際標準と世界情勢

Miyake Kazumasa
(miyake-kazumasa@sei.co.jp)
Sumitomo Electric Industries, Ltd.
Cyber Security R&D Department

2025.07. 23

SUMITOMO ELECTRIC GROUP

目次

1. 自己紹介
2. はじめに
3. 国際標準動向
4. 米国情勢
5. 英国情勢
6. 欧州情勢
7. 中国情勢
8. 国内情勢
9. まとめ
10. 付録
11. 参考文献

1. 自己紹介

住友電気工業株式会社　DX技術研究開発センター　サイバーセキュリティ研究開発部
三宅　和公（みやけ　かずまさ）miyake-kazumasa@sei.co.jp

【経歴】
現在：住友電工から産総研に出向・機械学習品質マネジメント(AIQM)に参加しAIセキュリティを担当。

●社会人になって以来、ブリッジ・ルータ・スイッチなどインターネットのインフラ機器の組み込みソフトウェアを開発。(全レイヤ：デバイスドライバ・OSから上位アプリまで)

●その後大学でGISシステム開発、映像リアルタイム配信・4K CATV関連システム、ASP開発を経てセキュリティ研究開発へ。
CATV事業者向けASP開発では設計・開発・保守・運用をしながらセキュリティの最前線で24H365Dの防衛実務に当たる。

●2010年頃から非公式統計学・データサイエンス勉強会を主宰。社会実装に向けた足掛かりとしてNEDO「実データで学ぶ人工知能講座」に2期連続参加して研鑽を積み、セキュリティ技術と組み合わせてAIセキュリティの研究開発へ。

そして2020年からAIQMに参加、LLM勉強会には2025年2月より参加。
AIセーフティ・AIセキュリティに必須の国際標準関連作業に参加。

2. はじめに-1

皆さんが日々開発している巨大言語モデル(LLM)・LLMを用いたアプリケーションは、やがて社会実装され、さまざまな方がさまざまな目的で利用する。

社会実装するには、製品やサービスを投入する社会・利用する人々を意識しなければならない。そのバックグラウンドには関係する世界動向があり、対応が求められる。

また、製品やサービスがどこで・誰に対して展開されるかによって認証を受けなければならない。認証を受けるには国際標準を把握しておく必要がある。（仮に認証を受けなくても従うべきデファクト国際標準があるなら対応していなければ市場は受け入れない。セキュリティや機能安全がよい例。）

★国際標準の意識は研究開発の最初から。後づけは難しい場合も（セキュリティや機能安全）

そこで、 AI（もちろんLLMを含む）セーフティ・AIマネジメントについて注目しておくべき国際標準、世界動向として米国・英国・欧州・中国・我が国の情勢のポイントを紹介する。

ここでは事実を端的に示す。

皆さんが既にご存じの事柄についてはご容赦を。

2. はじめに-2

ここでは、ポイントになる事柄に絞って紹介する。

●AIセキュリティ：AIを搭載するシステムに対する攻撃によって被害が発生する事象

●AIセーフティ：AIシステムの安全性確保

→様々な要因（例：機能安全・セキュリティ・プライバシーなど）によって被害が発生する事象を対象とする。つまり、AIセキュリティを包含する。

●AIマネジメント：AIシステムを安全・公平・信頼できる形で組織的に設計・運用・監視する活動（AIのリスク・影響評価・継続的改善）

→セーフティ・公平性・プライバシー・説明可能性・社会的影響・リスク評価/記録/監査を含む）

ハイライトのうち、

水色：各章のポイント

緑色：各章でポイントとなる用語

3. 国際標準化動向-1

★ISO/IEC 22989[1]を基礎概念、AIマネジメント全体としてISO/IEC 42001[2]があり、認証はISO/IEC 42006[4]でその中身はISO/IEC 42005[3]やISO/IEC 23894[5]が取り決めており、補足する文書が登場する予定である。

【AIマネジメント関連国際標準】

ISO/IEC 22989 – Artificial Intelligence Concepts and Terminology[1]

用語の定義、AIシステムの概念を定義（構成要素・ライフサイクル・学習能力や説明可能性などAIに伴う特性）

ISO/IEC 42001 — AI Management Systems[2]

●AIの統合的なリスク・影響評価・継続的改善を含む、AIマネジメントシステム構築の要求と枠組みを提供

●リスク関連では、特にClauses 6.1.2-6.1.4でリスク評価・AI影響評価・リスク処置の要件を定めており、ISO 23894などと連結して運用する構成になっている

ISO/IEC 42005 – AI System Impact Assessment[3]

AIシステムが人や社会に与える影響を「設計→運用→モニタリング」に至るライフサイクル全体で評価・アセスメントする方法を提供

ISO/IEC 42006 – Requirements for Certification Bodies[4]

ISO/IEC 42001 に基づくAIマネジメントシステムを審査・認証する第三者機関（認証機関）が満たすべき能力・手順・信頼性を規定

3. 国際標準化動向-2

★AIセキュリティなど個別の分野についても国際標準化が進捗しており、開発する対象に応じて利用する。

【テーマ個別の国際標準】

ISO/IEC 23894 — Guidance on Risk Management[5]
AI固有のリスク（バイアス、プライバシー、セキュリティなど）をライフサイクル全体で識別・評価・緩和・モニタリングするための包括的ガイダンスを提供

ISO/IEC DIS 27090 — Guidance for addressing security threats to artificial intelligence systems[6]
AIを狙うセキュリティ上の脅威（攻撃事例）を紹介し、AIを搭載するシステムのセキュリティを考えるうえでポイントとなる概念を紹介する。乞うご期待。

ISO/IEC CD 27091 - Privacy Protection[7]
AIシステムにおいてセキュリティから見たプライバシーの概念、リスクなどを紹介する。乞うご期待。

ISO/IEC 25059 — Systems and software Quality Requirements and Evaluation (SQuaRE)[8]
AIシステムにソフトウェア品質要求評価基盤（SQuaRE）を適用する方法

（出典：各ISO標準）

3. 国際標準化動向-3

【AIマネジメント関連の国際標準とテーマ個別の国際標準の見取り図】

凡例～黒字：発効済、青字：策定中

AIマネジメントに関連する国際標準

AIマネジメント国際標準

- ISO/IEC 42005 AI System Impact Assessment
- ISO/IEC 42006 AI Certification Bodies
- ISO/IEC 42001 AI Management System
- ISO/IEC 22989 AI Concepts and Terminology

テーマ個別の各種国際標準

- ISO/IEC CD 27091 AI Cybersecurity and Privacy
- ISO/IEC DIS 27090 AI Guidance for security threats
- ISO/IEC 23894 AI Risk Management
- ISO/IEC 25059 Quality system for AI

3. 国際標準化動向-4

【国際標準と各国情勢の関係】

●AIマネジメントに関連する国際標準を下地として米英欧中日の各国情勢が形成されている。

●中国についてはAIセーフティに関する法規制が登場しつつあるが、国際標準をどの程度意識しているかは不明なところがある。（内容としてはAI actを意識していると思われる。）凡例～黒字：発効済、青字：策定中、緑字：内容順次発行

米国

●産業促進が目標 EO 14179を中心に省庁や企業に任せる。

- M-25-21 Accelerating…
- M-25-22 Driving…
- EO 14179 トランプ流AI大統領令

英国

●段階的規制でイノベーションと安全性確保を両立

- UK AISI Research Agenda 今後の研究テーマ一覧
- A pro-innovation approach to AI regulation
- AI Regulation Bill

欧州

●AI actを中心に法規制を課す。LLMはGPAIに位置づけ具体策は整合規格を指定

- Data Act
- AI Act
- CRA(Cyber Resilience Act)
- DSA(Digital Services Act)

日本

●厳しい法規制は敷かずガイドライン類で事業者を支援

- 機械学習品質マネジメントガイドライン
- AIセーフティに関する評価観点ガイド
- AIセーフティに関するレッドチーミング手法ガイド
- AI事業者ガイドライン
- AI原則実践のためのガバナンス・ガイドライン

中国

●AI actを参考。違いは、国家安全・国家統制が強く優先（不明な点も多い）

- 生成AIサービスの管理に関する暫定措置
- サイバーセキュリティ法
- データセキュリティ法
- 個人情報保護法

AIマネジメントに関連する国際標準

- ISO/IEC 42005 AI System Impact Assessment
- ISO/IEC 42006 AI Certification Bodies
- ISO/IEC 42001 AI Management System
- ISO/IEC 22989 AI Concepts and Terminology
- ISO/IEC 23894 AI Risk Management
- ISO/IEC 25059 Quality system for AI
- ISO/IEC DIS 27090 AI Guidance for security threats
- ISO/IEC CD 27091 AI Cybersecurity and Privacy

3. 国際標準化動向-5

★AIマネジメントの国際標準ISO/IEC 42000シリーズはセキュリティの国際標準ISO/IEC 27000シリーズを参考に開発された。

★ISO/IEC 27000シリーズは、
ISO/IEC 27000[9]：用語の定義と概念
ISO/IEC 27001[10]：ISO/IEC 27000シリーズ全体の概念を示す
ISO/IEC 27002[11]：管理策例
ISO/IEC 27005[12]：リスクアセスメント（具体的なセキュリティの分析・対策の進め方が定義されている。

★これらのうち、ISO/IEC 27005[12] に当たるもの（具体的に何をしなければならないかが分かるもの）がISO/IEC 42000シリーズには存在しない。→ISO/IEC 42000シリーズの課題

★ISO/IEC 27005[12]は、ISO/IEC 27000シリーズ策定の中でも難航を極めた。

★ISO/IEC 42000シリーズにおいて、ISO/IEC 27005[12]に対応するものが望まれる。ISO/IEC 42001[2],42005[3],42006[4]だけでは具体的に「何をすればよいか」が表せない。
→現状としては案件ごとに課題を抽出して対策する。

（出典：各ISO標準）

4. 米国情勢-1

★トランプ政権となり一律な安全性担保から産業促進にシフト[1]。ただし、安全性に向けた仕組みや文書は継承され、安全性は各省庁の機関に任せる。[2]

バイデン政権下：大統領令EO 14110[13]配下にNIST AI RMF[14]やPlaybook[15]が制定され省庁横断の仕組みを構築しようとしていた。AIセーフティ・AIリスクに向けた取り組みはUS AISIが担っていた。

トランプ政権下：EO 14110[13]の「置き換え」としてEO 14179[16]が発出された。US AISI(Safety Institute)は廃止された訳ではないがCAISI(Standards and Inovation)に改名、AIリスクに対する対策のほかAIを用いた産業振興・国家安全保障にも重点を置く。

●NISTは引き続き生成AIを含めた取り組みを推進、Elham Tobassiも健在。NIST AI RMFも存続・活動しており、廃止された訳ではない。

●NIST AI RMF[14]は省庁横断的な位置づけではなくなった。各省庁にはCAIOを配置し、AIセーフティ・AIリスクを担う部局を設置し、各省庁によりフィットした対応を行う方針とした。

1 大統領令EO 14179[16]関連文書M-25-21[17]を参照。
2大統領令EO 14179[16]関連文書M-25-21[17]を参照。

4. 米国情勢-2

●EO 14179[16]がEO 14110[13]の「置き換え」として制定され、配下にM-25-21[17]、M-25-22[18]が制定された。各所轄部局は説明責任・評価の義務を負う。EO 14179[16]関連文書M-25-21[17]では

・“High Impact AI”を定義し、セーフティ取り組み強化、プロセスの確立を求める。～人権・安全・重要インフラ・戦略資産など広範囲の及ぶリスク源であり、その出力（AIが出す結果や判断）が意思決定や行動の「主要な根拠」となるAIを指す。以下は引用。

【意思決定や行動が与える影響】そのAIを使うとどのような影響があるか

- 法的な影響がある（例：権利の剥奪、義務の発生）
- 実質的な影響がある（例：経済的損失やサービス提供の可否）
- 拘束力がある（例：裁定、認可、不認可）
- 社会的・制度的に重要な影響がある（例：安全、健康、自由などに関わる）

【High Impact AIが何に関する出力をするか】そのAIがどのような出力をするか。

- 個人または団体の市民権、市民的自由、プライバシー
- 個人または団体の教育、住宅、保険、信用、雇用など
- 人間の健康と安全；
- 個人または団体による、重要な政府資源またはサービスへのアクセス
- 重要インフラまたは公共の安全
- 戦略的資産または資源（高価値の財産、機微または機密とされた情報を含む）

・High Impact AIには安全性・セキュリティ・回復力を求める。充足できない場合は使用を中止する。

・セーフティにおいては、重大事故・DeepFakeを使った国家の操作・敵対国によるAIの兵器化）も対象：国家安全保障に向け取り組み強化

・モニタリングによりAI機能のパフォーマンスとセキュリティに対する悪影響（Adverse Impacts）を特定

●AIリスクの整理として、MIT AI Risk Repository[19]も参考にするとよい。

●新しい情報が登場した場合は要チェック。

4. 米国情勢-3

★以下の事例のように企業はリスクを恐れてリスクの検出・評価・対応について調査・検討を進めている。（OpenAI、Anthropicなど）

【企業の取り組み例】

(1) Open AI

●Safety evaluations hub [20](https://openai.com/safety/evaluations-hub/#jailbreak-evaluations)

モデルのセーフティベンチマーク結果をまとめて公開（反応してはならないコンテンツへの対応・ジェイルブレイク・ハルシネーションなど）

●Open AI Model Spec[21] (https://model-spec.openai.com/2025-04-11.html)

ゴールのずれ、誤った出力、有害な指示への対応など、セーフティ評価の指針を整理しまとめる。

(2) Anthropic

●Activating AI Safety Level 3 Protections[22] (https://www.anthropic.com/news/activating-asl3-protections)

Jailbreak、重みパラメータの盗難、アクセス制御、権限昇格を取り上げる。

5. 英国情勢-1

★EU AI actを意識しつつ、規制は段階的に進めイノベーション推進と安全性確保の両立を目指す。規制ではなく「原則」。特に生成AIのリスクについては具体的な取り組みが進む。

★ “A pro-innovation approach to AI regulation” [23]やAI Regulation Bill[24]を原点としてResearch Agenda[25]を策定し、これまでにないAIのリスクに向けた取り組みを開始

●”A pro-innovation approach to AI regulation “[23]：法律でがっちり縛るのではなく「原則」を設ける。

・原則は、Safety・Transparency・Fairness・Accountability・Contestability（説明責任と是正可能性）

・これらを分野別に実現するアプローチを持つ（医療・金融など）

・制度的基盤整備：AI Regulatory Sandbox（AI規制のサンドボックス）・AI Risk Function（AIリスク分析機能）・AI Security Institute（AISI）

●AI Regulation Bill[24]によって法で5原則を法制度化しようとする動きもある。ただし法制化は不透明。

●UK AISI Research Agenda[25]の登場：生成AIを意識した内容

・リスク6例：Cyber Misuse(悪用)・Dual-use Science risks（科学におけるAI活用と悪用）・Criminal Misuse(犯罪に向けた悪用)・Autonomous Systems risks(自律によるリスク)・Societal Resilience risks(AIによってもたらされる混乱への耐性)・Human Influence risks（AIによる心理的操作、説得、人間の意思決定への介入）

・評価手法：Inspectの開発と提供（後述）、階層ベイズモデル（HiBayES）によるモデル性能の評価、エージェント評価（自律的タスク遂行・外部ツール呼び出し・自己複製）

5. 英国情勢-2

★生成AIにおいて、モデルが知っている事実とは異なる出力をする場合や性能を低く見せようとする挙動、マルチエージェント化したときに現れる創発性などのリスクに対する懸念があり、Research Agenda[25]では懸念するリスクやベンチマーク開発を紹介している。

【Research Agenda[25]に登場する、生成AI固有と思われるリスクや懸念に向けたアプローチ】

5.1. リスクに関連するもの

●Constitutional Classifier[28]を紹介(Constitutionを与えることでjailbreakを検出・防御)

●Sandbagging：AIが外部から自らの性能が低く見えるようにする挙動を懸念し評価する。

●SWE-bench[26]：AIモデルやエージェントがサイバーセキュリティ関連のタスク（例：脆弱性検出、マルウェア分類、攻撃予測など）をどの程度こなせるかを評価するためのベンチマークする。

●マルチエージェントシステムを分析することで、シングルエージェントでは現れない創発的な動作(emergent behaviour)を理解できるようになる。

●Agent Hijackを取り上げる。ゴール設定の改ざんが主であり、技術的な手法が存在する（ただし研究レベル）。

●ControlArena[29]を提供：エージェント型AI（特にコード実行エージェント）を対象に、実際のフロンティアAIラボ環境に近い複雑さを持つ設定で、「制御（control）」の評価を行うフレームワーク。
「制御（Control）」とは、単なる「命令どおりに動かす」ことではなく、高度な・自律的なAIエージェントが望ましくない行動（特に悪意ある行動）をとらないようにするための、全体的な仕組みや戦略。

6. 欧州情勢-1

★AI act[32]を中心。AIのアプリケーションのうち、禁止するもの・高いリスクが想定されるものを例示し事業者に対する要求など規制を示す。

★LLMはGPAIに位置づけられ、システミックリスク（大規模で社会的影響が大きい基盤モデルやGPAIのリスク）に対する追加事項への対応も求める。

ここでは欧州情勢をAI act[32]を軸に紹介しよう。

【AI actの[32]概要】

●AIリスクを「システムによる安全性・基本権・法的義務の侵害の可能性」と捉え、リスクベースアプローチを導入

- Unacceptable Risk（許容できないリスク Article 5）：禁止
- High Risk（高リスク Article6/Annex III）：医療、交通、インフラ、雇用などに対し、厳格な要件を規定（リスク管理・説明責任・記録保持など）Article 15参照

このほか、Limited Risk（限定リスク）、 Minimal Risk（最小リスク）を定義

●AIのセーフティは「人間の安全・健康・環境・システムの信頼性を保つこと」に重点

・Article 15：ロバストネスとセキュリティの要件、データセットにおける統計的偏りの排除

・Article 14：Human Oversight（監視）の要件

6. 欧州情勢-2

●AIのマネジメントとしてQuality Management Systemを求める。

設計監理・設計検証・品質管理・検証手順・技術仕様・データ管理手順・リスク管理システム・運用モニタリングシステム・重大インシデント報告手順に対して文書化を求める。

●LLMは、AI act[32]の中ではGPAI(General Purpose AI)として扱われる。技術文書の作成と提出・著作権法の順守・訓練データの要約と開示・透明性の確保・これらの文書化を求める。

・モデルの重みや構造、利用情報が一般公開されているGPAIモデルについては、上記の一部義務が適用除外となる場合も。

・システミックリスクに対しては、義務が追加される。

モデル評価・敵対的テスト、リスク軽減措置、インシデント報告、セキュリティ強化

システミックリスクの事例：定義から以下のようなものが考えられるだろう。

偽情報の大量拡散、偽の音声動画の大量生成、マルウェアコード生成、大規模な公平性侵害など。

・GPAIに関する規制は2025年8月2日から段階的に施行され、違反時の罰則は、最大でグローバル売上高の7%または3,500万ユーロ（約55～60億円）のいずれか高い方

・欧州AI委員会、科学委員会、諮問フォーラムなどの諮問機関が設置され、GPAIのリスク評価や規制運用を監督

6. 欧州情勢-3

【AI act以外の注意点】

●AIマネジメント

・AI HLEG：欧州委員会が2018年に設置した諮問機関であり、AIに関する倫理・法制度・技術の課題に対して助言してきた。

・AI HLEGは、”Ethics guidelines for trustworthy AI “[33]で「信頼できるAIの7原則」（Human Agency and Oversight・Technical Robustness and Safety・Privacy and Data Governance・Transparency・Diversity, Non-discrimination and Fairness・Societal and Environmental Well-being）を示す。

●AIガバナンス

ALTAI（Assessment List for Trustworthy AI）：上記の”Ethics guidelines for trustworthy AI “[33]を実務的に活用するために作成した自己評価ツール、AI HLEG開発。

●AIのガバナンスとして

・European AI Office設置：EU域全体での実施監督・執行・国際連携を担う中央機関

GenAI4EU：生成AI（Generative AI） の開発・導入を支援し、「欧州発の人間中心かつ信頼性の高いAI」を促進する戦略的イニシアティブ、モデル開発から社会実装まで目指す。

・EU域全体での実施監督・執行・国際連携を担う中央機関設置予定

7. 中国情勢-1

（本書では、法規制については、各法規制の後に記載した仮の和名で呼ぶ。）

★法規制はAI actを参考にすべきとの見解あり。[3]

★違いは、消費者保護や個人の権利の保護よりも国家安全・国家統制が強く優先され、社会秩序攪乱罪・国家分裂罪・虚偽情報散布罪設けられている。[4]

ここでは方向性が分かる2つの法規制を紹介しよう。

7.1.生成式人工智能服務管理暫行辦法[35]（生成AIサービスの管理に関する暫定措置）

- ●中国初の生成AIに関する規制
- ●中华人民共和国网络安全法[37]（中華人民共和国サイバーセキュリティ法）、中华人民共和国数据安全法[38]（中華人民共和国データセキュリティ法）、中华人民共和国个人信息保护法[39]（中華人民共和国個人情報保護法）の遵守を求める。
- ●知的財産権の侵害を禁止する条項がある。
- ●誰が・いつ・どのような目的で・何に使うのかの監視を求める。

3 欧盟人工智能政策分析（"Analysis of EU Artificial Intelligence Policy"）[34] 参照
4生成式人工智能服務管理暫行辦法[35]（生成AIサービスの管理に関する暫定措置）、互聯網信息服務深度合成管理規定[36]（ディープフェイク規制）参照

7. 中国情勢-2

（本書では、法規制については、各法規制の後に記載した仮の和名で呼ぶ。）

特徴をよく表している条文（和訳）を紹介する。

第4条「社会主義の核心的価値観を堅持し、国家権力の転覆、社会主義体制の転覆、国家の安全と利益の危機、家庭の損害、国家の団結と社会の安定を損なう扇動を発生させ、テロリズム、過激主義を促進し、民族憎悪、民族差別、暴力、わいせつ、ポルノを促進し、虚偽で有害な禁止内容を発生させてはならない」

第8条 データラベリングのための具体的で運用可能なラベリング品質評価、サンプリング方式でラベリングされた内容の正確性の検証、ラベリングの仕様について監督・指導を求める。

第17条 「世論属性または社会動員能力を持つ生成的人工知能サービスを提供し、国家安全評価の関連規定に従って実施されるべきであり、記録のためのアルゴリズムを実行するために、アルゴリズムの管理規定のアルゴリズム勧告、および変更、廃止の手続きを行うものとする。」

第19条「国家機密、商業秘密、個人的な法律の知識の中で自分の職務を遂行する上で、関連機関や人員の生成的人工知能サービスの安全性評価と監督検査に参加し、秘密を保持しなければならない、漏洩したり、違法に他人に提供してはならない。」

出典：生成式人工智能服務管理暫行辦法[35]（生成AIサービスの管理に関する暫定措置）

7. 中国情勢-3

（本書では、法規制については、各法規制の後に記載した仮の和名で呼ぶ。）

互聯網信息服務深度合成管理規定[36]（ディープフェイク規制）

- ●サービス提供者・開発者・利用者それぞれに厳しい規制を敷く。
- ●罰金・量刑の規則がある。
- ●中华人民共和国网络安全法[37]（中華人民共和国ネットワーク安全法）、中华人民共和国数据安全法[38]（中華人民共和国データ安全法）、中华人民共和国个人信息保护法[39]（中華人民共和国個人情報保護法）、互联网信息服务管理办法[40]（インターネット情報サービス管理弁法）の遵守を求める。
- ●国家安全、国家のイメージ、国家の利益または社会公共の利益に関わる特殊な物体、シーン等の非生体認証情報を生成または編集する機能を有するモデル・ツール類に対して安全性評価を用いる。

第9条「深層合成サービス利用者の実在する身元情報を認証しなければならず、実在する身元情報を認証していない深層合成サービス利用者に対して、情報提供サービスを提供してはならない。」

第10条「深層合成サービス提供者は、深層合成コンテンツの管理を強化し、技術的または手動的な方法を採用して、深層合成サービス利用者の入力データと合成結果を監査しなければならない。」

8. 国内情勢-1

★厳しい法規制は敷かず社会実装に向けたガイドライン類で事業者を支援。省庁などからガイドライン類が発出されている。特にAI事業者ガイドラインに注目。[5]

1. AIリスク

経産省「AI事業者ガイドライン（2024）」[41]

●AIの開発・提供・利活用におけるリスクベース管理の推進ために著された。

●共通の指針（バリューチェーン全体で取り組むテーマ）：人間中心、安全性、公平性、プライバシー保護、セキュリティ確保、透明性、アカウンタビリティ、教育・リテラシー、公正競争確保、イノベーション

●高度なAIシステムに関係する事業者に共通な指針：リスク・インシデントと脆弱性のパターンを特定、透明性の確保、情報共有とインシデント報告、リスク管理とガバナンス方針の策定など

●ガバナンスの構築：アジャイルガバナンス（ゴール設定・システムデザイン・運用・評価・リスク分析）

●AI開発者・AI提供者・利用者に対し重要事項を示す：AI開発者には広島AIプロセスについて追加事項

●開発者・提供者・利用者それぞれに取り組み事項を追加

●説明責任・記録保持・苦情対応の義務化を求める。

●出力結果のモニタリングとトレーサビリティ強化

●技術的に具体的な検討内容については「機械学習品質マネジメントガイドライン第4版」を引用。

5 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html

8. 国内情勢-2

2. AIセーフティ

AI Safety Institute Japan（AISI）：2023年発足、IPA内に設置、経産省が主導

AIシステムの安全性・信頼性・公平性・セキュリティなどの観点から「AIセーフティ評価」を実行・指導し、国際的な整合性・標準化を目指す。

●AIセーフティに関する評価観点ガイド[42]

（https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.00_ja.pdf）

LLMをはじめとするAIシステムの安全性評価について、11の評価観点（有害情報の出力制御、偽誤情報の出力・誘導の防止、公平性と包摂性、ハイリスク利用・目的外利用への対処、プライバシー保護、セキュリティ保護、説明可能性、ロバスト性、データ品質、検証可能性）の項目を体系化。→項目を炙り出すための概念の整理、具体的な項目や技術的な中身の話はない。

●AIセーフティに関するレッドチーミング手法ガイド[43]

（https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_ja.pdf）

レッドチーミングテストを想定し、重要な脆弱性やセキュリティギャップを検出するための手順段階的アプローチを解説した 。手順を示したものなので、具体的なテストの観点は、OWASP top 10 for LLMやAgentic AI - Threats and Mitigationsを参照する必要がある。

●国際連携：米英シンガポールなど、広島プロセスの技術支援

●マンスリーレポート

8. 国内情勢-3

2. AIセーフティ（つづき）

AISI事業実証ワーキンググループ・ビジョンペーパー[44]

(https://www.ipa.go.jp/pressrelease/2025/press20250711.html)

●事業個別（ヘルスケア・ロボティクスなど）に評価シナリオ・評価データセット・評価ツールについてガイドするSWG（※1）を設立

●適合性評価・データ品質など各種事業に共通する分野についてもガイドするSWG（※2）を設立

●※1→※2：現場目線のニーズを提供、※2→※1：現場で使う標準的な評価指針

8. 国内情勢-4

3. AIマネジメント・ガバナンス

3.1.経産省「AI原則実践のためのガバナンス・ガイドライン」[45]

●AIガバナンスの工程それぞれに対して「行動目標」と行動目標の実施例を示す。

●工程は、工程・リスク分析、ゴール設定、システムデザイン（AIマネジメントシステムの構築）、運用、評価であり、これらをアジャイルガバナンスで回す。それぞれにガイドを示す。

●ログ管理、再学習管理、異常検知についても述べる。

3.2.ISO/IEC 42001[2]（AIマネジメントシステム）との連携

産総研・AISIにより、国内指針と国際規格との橋渡しを図る。

4.AI品質マネジメント

●産総研：機械学習品質マネジメントガイドライン[46]、機械学習品質マネジメントイニシアティブ

●QA4AI：AIプロダクト品質保証ガイドライン[47]

9. まとめ

まとめとして最も規制の強い欧州AI actとの関係を軸に整理してみよう。

国	欧州委員会AI act	米国	英国	日本	中国
状況(AI Actとの違い)	●厳格な法規制、厳しい罰則・罰金 ●高リスクAIについては厳格な規制 ●EUとしてAIリスクに取り組む ●具体的な実施内容は整合規格へ	●規制よりも産業振興 ●AIリスクは各省庁・各企業それぞれで管理、高リスクへの対応を求める。	●規制ではなく「原則」を設け厳しく規制しない ●段階的に規制 ●政府としてAIリスクに取り組む	●規制ではなくガイドラインで分析方法や対応方法を提示 ●政府としてAIリスクに取り組む ●42000シリーズへの積極的な取り組み	●規制は設けるが安全や個人の権利の保護よりも国家統制を目標、厳しい罰則 ●国家統制のためのAIリスク管理
AIセーフティ組織	AI Office	AISI→CAISI	AISI UK	AISI Japan	組織開設予定

10. 付録-1

欧州AI関連法規のリリースタイムライン

		2022	2023	2024	2025	2026	2027	2028	2029	2030
AI act	発効（ただし各種要件は順次適用）				▲8月					
	特定AIシステムの禁止とAIリテラシーに関する要件適用				▲2月					
	GPAI提供者の行動規範展開				▲5月					
	規則（通知機関・GPAIモデル・ガバナンス・守秘義務・罰則）の適用開始→猶予3年				▲8月 →			▲8月猶予期間（3年）		
	高リスクAI義務、適用開始→猶予6年					▲8月 →				
Data act	発効			▲1月						月猶予期間（6年：2032年８月まで）
	施行				▲9月12日					
CRA	欧州理事会がCRAについて合意		▲7月							
	欧州議会と理事会がCRAについて政治的合意			▲12月						
	欧州議会がCRAを承認			▲3月						
	CRA発行				▲12月10日					
	脆弱性・インシデント報告義務の適用					▲9月10日				
	CRAの全面適用および整合規格の公表							▲12月11日		
DSA	透明性のある報告やVLOPに対する要件など一部適用開始	▲11月16日								
	オンラインプラットフォームやオンライン検索エンジンは月間アクティブユーザ数を公表		▲2月17日							
	全面施行、加盟国はDigital Services Coordnatorsを指名			▲2月17日						

特に注意が必要な法規制

AI act[32]

Data act[48]

CRA(Cyber Resilience Act)[49]

DSA(Digital Services Act)[50]

11. 参考文献-1

[1] ISO/IEC 22989:2022 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology

[2] ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system

[3] ISO/IEC 42005:2025 Information technology – Artificial intelligence (AI) – AI system impact assessment

[4] ISO/IEC 42006:2025 Information technology – Artificial intelligence – Requirements for bodies providing audit and certification of artificial intelligence management systems

[5] ISO/IEC 23894:2023 Information technology – Artificial intelligence – Guidance on risk management

[6] ISO/IEC DIS 27090 Cybersecurity – Artificial Intelligence – Guidance for addressing security threats to artificial intelligence systems

[7] ISO/IEC CD 27091 Cybersecurity and Privacy – Artificial Intelligence – Privacy protection

[8] ISO/IEC 25059:2023 Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems

[9] ISO/IEC 27000:2018 Information technology – Security techniques – Information security management systems – Overview and vocabulary

[10] ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection – Information security management systems – Requirements

[11] ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection – Information security controls

[12] ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection – Guidance on managing information security risks

[13] United States, Executive Office of the President [Joe Biden]. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 30 Oct. 2023.

https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[14] National Institute of Standards and Technology (United States of America). NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023.

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

11. 参考文献-2

[15] National Institute of Standards and Technology (United States of America). NIST AI RMF Playbook, May 2024.

https://airc.nist.gov/airmf-resources/playbook/

[16] United States, Executive Office of the President [Donald Tramp]. REMOVING BARRIERS TO AMERICAN LEADERSHIP IN ARTIFICIAL INTELLIGENCE, 23 January 2025.

https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[17] United States, Executive Office of the President Office of Management and Budget. M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust, 3 Apr. 2025.

https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

[18] United States, Executive Office of the President Office of Management and Budget. M-25-22: Driving efficient acquisition of artificial intelligence in government, 3 April 2025.

https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf

[19] Massachusetts Institute of Technology. The AI Risk Repository, March 2025.

https://airisk.mit.edu/

[20] OpenAI. Safety evaluations hub, 14 May 2025.

https://openai.com/safety/evaluations-hub/

[21] OpenAI. Open AI Model Spec, 11 April 2025.

https://model-spec.openai.com/2025-04-11.html

[22] Anthropic. Activating AI Safety Level 3 Protections, 23 May 2025.

https://www.anthropic.com/news/activating-asl3-protections

11. 参考文献-3

[23] UK Government, Department for Science, Innovation and Technology, A Pro-Innovation Approach to AI Regulation, 3 August 2023.

https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[24] UK Parliament. Artificial Intelligence (Regulation) Bill [HL] (As Introduced), 4 Mar. 2025.

https://bills.parliament.uk/publications/59353/documents/6128

[25] UK AI Safety Institute, Department for Science, Innovation and Technology, UK AI Safety Institute Research Agenda, 2023.

https://www.aisi.gov.uk/research-agenda

[26] Kocetkov, Dmitry, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? arXiv preprint arXiv:2310.06770 (2023). https://arxiv.org/abs/2310.06770.

[27] UK AI Safety Institute. Inspect, 2024.

https://inspect.aisi.org.uk/

[28]Sharma, Tong, et al. “Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming.” arXiv preprint arXiv:2501.18837 (2025). https://arxiv.org/abs/2501.18837.

[29] UK AI Safety Institute. ControlArena, 2025.

https://github.com/UKGovernmentBEIS/control-arena

[30] Khattab, Singhvi, et al. “DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines” arXiv preprint arXiv:2310.03714 (2023). https://arxiv.org/abs/2310.03714

[31] Hu, Lu, et al. ”Automated Design of Agentic Systems” arXiv preprint arXiv:2408.08435v2 (2024). https://arxiv.org/abs/2408.08435

[32] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union, L 202, 13 June 2024, pp. 1-266. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

11. 参考文献-4

[33] European Commission, High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. April 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[34] Li, Feng, et al. 欧盟人工智能政策分析（Analysis of EU Artificial Intelligence Policy), 2025

https://cn.erytis.com/index.php/Cs/article/view/43

[35] Cyberspace Administration of China. Interim Measures for the Management of Generative Artificial Intelligence Services（生成式人工智能服務管理暫行辦法). 13 July 2023, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[36] Cyberspace Administration of China. Provisions on the Administration of Deep Synthesis of Internet Information Services（互联网信息服务深度合成管理规定). 11 Dec. 2022, https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm.

[37] Standing Committee of the National People’s Congress. Cybersecurity Law of the People’s Republic of China（中华人民共和国网络安全法). 7 Nov. 2016, https://www.gov.cn/xinwen/2016-11/07/content_5129723.htm.

[38] Ministry of Information Industry. Measures for the Administration of Internet Information Services（互联网信息服务管理办法). 25 Sept. 2000, https://www.gov.cn/gongbao/content/2000/content_60531.htm

[39] Standing Committee of the National People’s Congress. Personal Information Protection Law of the People’s Republic of China（中华人民共和国个人信息保护法). 20 Aug. 2021, https://www.gov.cn/xinwen/2021-08/20/content_5632486.htm.

[40]互联网信息服务管理办法

Ministry of Information Industry. Measures for the Administration of Internet Information Services（互联网信息服务管理办法). 25 Sept. 2000, https://www.gov.cn/gongbao/content/2000/content_60531.htm.

[41] 経済産業省 AI事業者ガイドライン（第1.1版）2025年3月. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_1.pdf

[42] AIセーフティインスティテュート AIセーフティに関する評価観点ガイド 2025年3月. https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.10_ja.pdf

[43] AIセーフティインスティテュート AIセーフティに関するレッドチーミング手法ガイド（第1.10版） 2025年3月 https://aisi.go.jp/assets/pdf/J1_ai_safety_RT_v1.10_ja.pdf

[44] AIセーフティインスティテュート AISI事業実証ワーキンググループ・ビジョンペーパー 2025年7月 https://aisi.go.jp/assets/pdf/Vision_paper_v1.01.pdf

11. 参考文献-5

[45] 経済産業省 AI 原則の実践の在り方に関する検討会 AIガバナンス・ガイドライン（Ver.1.1） 2022年1月 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_1.pdf

[46] 産業総合技術研究所 機械学習品質マネジメントガイドライン第4版 2023年12月 https://www.digiarc.aist.go.jp/publication/aiqm/AIQuality-requirements-rev4.2.0.0113-signed.pdf

[47] Aiプロダクト品質保証コンソーシアム AIプロダクト品質保証ガイドライン2025.04版 https://www.qa4ai.jp/download/

[48] European Parliament and Council of the European Union. Regulation (EU) 2023/2854 of 13 December 2023 on harmonised rules on fair access to and use of data (Data Act). Official Journal of the European Union, L, 22 Dec. 2023, pp. 1-85. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202302854.

[49] European Parliament and Council of the European Union. Regulation (EU) 2024/1252 of 11 April 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020 (Cyber Resilience Act). Official Journal of the European Union, L 202, 23 October 2024, pp. 1-159. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202402847.

[50] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, L 277, 27 Oct. 2022, pp. 1-102. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065.

SUMITOMO ELECTRIC

Connect with Innovation

https://sumitomoelectric.com/

SUMITOMO ELECTRIC GROUP