AIと消費者保護：EUのAI関連法制の観点から

1．AI規制論の焦点：生成AI以前と以降
2．EU AI法案：生成AI以前と以降
3．EUプラットフォーム規制と（生成）AI
4．いくつかの論点

2024年3月26日
生貝直人　博士（社会情報学）
一橋大学大学院法学研究科ビジネスロー専攻教授

1．AI規制論の焦点：生成AI以前と以降

- 生成AI以前（2016頃～）：情報を処理するAIと法
  - EU以外はソフトロー重視
  - 主な焦点リスクは**製品安全＋プロファイリング**

- 生成AI以降（2022終盤～）：情報を生成するAIと法
  - EU以外もハードローを意識
  - 主な焦点リスクは**偽・誤情報と情報環境全体への影響**
  - 関連して国家（経済）安全保障や、競争政策の前面化

※本資料でのEU AI法案の内容は、特に言及が無い限り2024年3月13日欧州議会採択テキストに基づく。
https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

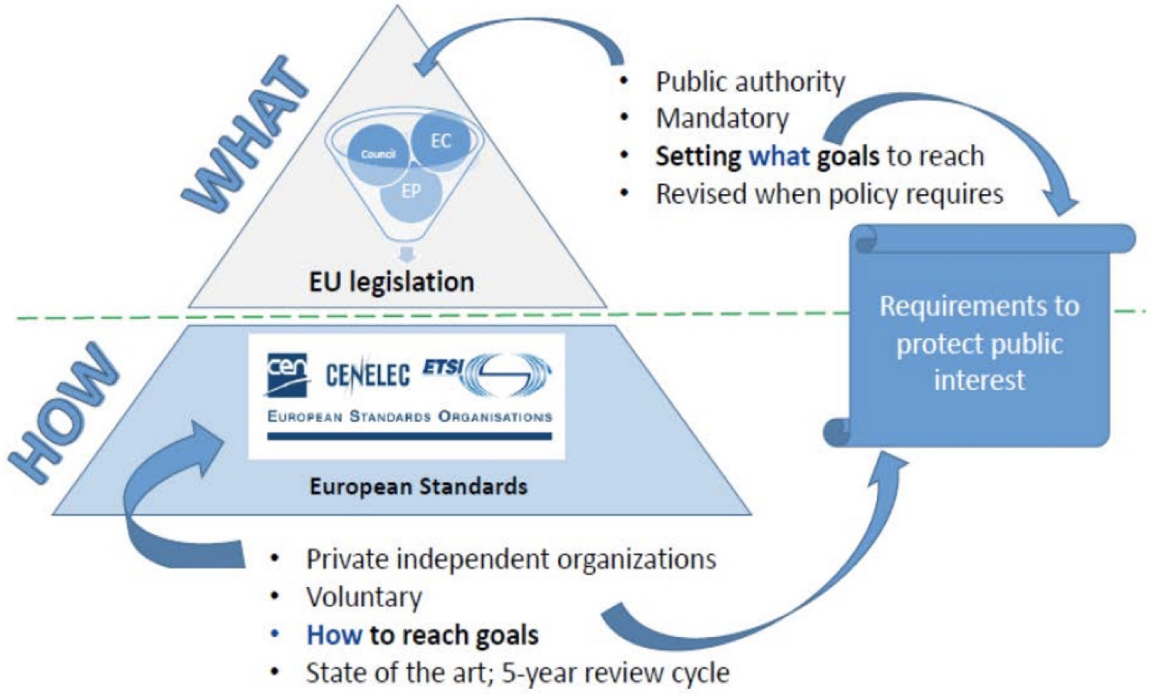
2．EU AI法案：生成AI以前

- AIシステムをリスクに応じて4段階に分類した規律を置く
  - 許容できないリスク（禁止されるAI利用行為）
  - **ハイリスク（適合性評価等の義務）**
  - 限定的リスク（透明性義務）
  - 低リスク・無リスク（拘束力の無い行動規範）
- 禁止されるAI慣行（一部）
  - 判断能力に著しい影響を与えるサブリミナル技法や操作的・欺瞞的技法
  - 年齢、障害、特定の社会的・経済的状況に起因する脆弱性の悪用
- ハイリスクAIのカテゴリー
  - 既存EU法での適合性評価義務対象：機械、玩具、レジャー用船舶、リフト、爆発性雰囲気装置、無線機器、圧力機器、索道設備、個人用保護具、ガス機器、医療機器
  - AI法での新たな指定：バイオメトリクス（遠隔生体識別・感情識別）、インフラ管理・運用、教育や職業訓練での学生や希望者の評価や受入れの合否、雇用、労働管理、自営業へのアクセス、重要な民間・公共サービス（公的支援金給付、融資、緊急対応措置）、法執行、移民・亡命・国境管理、司法又は民主主義プロセス

2．EU AI法案：生成AI以前

- ハイリスクAIシステムの要求事項
  - リスクマネジメントシステムの構築、データとデータガバナンス、技術文書、記録保持、透明性と利用者への情報提供、人間による監視、正確性・堅牢性・サイバーセキュリティ
- ハイリスクAIシステム提供者の義務
  - 上記要求事項の遵守確保、品質マネジメントシステムの構築、技術文書の作成、適合性評価の実施、自動生成ログの保存、本規則不適合時の是正、リスク検知時の監督機関への通知、監督機関との協力、CEマーク付与、データベース登録等
- ハイリスクAI配備者の義務
  - **基本権影響評価の実施**
- 整合規格（Harmonized standard）による要求事項の具体化
  - 欧州委の要請に応じて、CEN (欧州標準化委員会)やCENELEC (欧州電気標準化委員会)、ETSI (欧州通信規格協会)が発行する欧州統一規格

参考：AI法案と整合規格・共同規制



CEN-CENELEC “Drafting Harmonized Standards in support of the Artificial Intelligence Act (AIA)” より

「例えば、EU市場に投入される製品やサービスの認証に使用される整合規格も、共同規制手法（co-regulatory instruments）の一種である。この点、ECが発表したAI法の提案は、CEマークを取得しAIソフトウェアを市場に出すためのコンプライアンスツールとして整合規格に依存している。整合規格は「New legislative framework」（NLF）の一部であり、リスクの高い製品やサービスを特徴とする一部の技術・産業分野を規制するために採用されているアプローチである。NLFでは、法律が一般的な要求事項のみを定義し、どのように遵守するかは法律の受け手に委ねられる。」

Vigna Francesco [2022] “Co-regulation Approach for Governing Big Data: Thoughts on Data Protection Law”

「AIシステムの導入は、基本権に重要な規範的意味を持つ。標準の規制的役割を含む民主的な立法・規制プロセスは、技術の社会的影響と関連する利害関係者の参加を考慮すべきである。しかし、規制を（民間の）標準設定機関に委ねる主な問題は、標準化がしばしば功利主義的、技術主義的、非民主的と見なされることである。 例えば、利害関係者が標準設定に参加したとしても、議論への具体的な貢献が保証されるわけではない。」

Christopher Marsden “Artificial Intelligence Co-Regulation: The role of standards in the EU AI Act”

2．EU AI法案：生成AI以降

- **汎用目的AIモデル（生成AIはその典型）を含むコンテンツ生成AI提供者の義務**
  - 出力が機械可読形式でマークされ、人為的に生成又は操作されたことを検知可能とする
- **汎用目的AIモデル提供者**の義務
  - 設計や学習等の技術文書作成と当局への提供
  - 下流事業者への情報開示
  - DSM著作権指令4条（学習データオプトアウト）遵守措置、学習データ要約
- **システミックリスクを有する汎用目的AIモデル（10^25FLOPs以上等）提供者**の義務
  - システミックリスク特定・軽減のためのレッドチームテスト実施・文書化を含むモデル評価
  - **システミックリスクの評価・軽減**
  - 重大インシデントへの対応文書化と当局への報告
  - サイバーセキュリティ対策

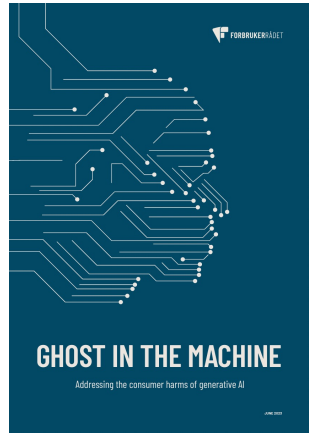
→それぞれの義務は整合規格により具体化、それまでは欧州委員会主導で策定する行動規範（codes of practice）の遵守

3条(65)「「システミックリスク」とは、汎用目的AIモデルの高インパクト能力に特有のリスクであって、その影響範囲の広さにより連合市場に重大な影響を及ぼし、または公衆衛生、安全、治安、基本権もしくは社会全体に対する実際の若しくは合理的に予見可能な悪影響により、バリューチェーン全体にわたって大規模に伝播し得るリスクをいう」

2．EU AI法案：生成AI以降

- 前文133「**さまざまなAIシステムが大量の合成コンテンツを生成できるようになり、人間が生成した本物のコンテンツとの区別がますます難しくなっている。**こうしたシステムが広く利用可能になり、その能力が高まることは、情報エコシステムの完全性と信頼性に重大な影響を及ぼし、**誤情報や大規模なマニピュレーション、詐欺、なりすまし、消費者への欺瞞といった新たなリスクを引き起こす。**こうした影響、技術の進歩の速さ、情報の出所を追跡するための新たな手法や技術の必要性を考慮すると、こうしたシステムのプロバイダーに対し、機械が読み取り可能な形式で表示し、その出力が人間ではなくAIシステムによって生成または操作されたことを検出できる技術的ソリューションを組み込むことを求めることが適切である。(…)」

"GHOST IN THE MACHINE: Addressing the consumer harms of generative AI" (Norwegian Consumer Council, June 2023)
「人間の行動をエミュレートする能力に制限を設けることなく、生成AIモデルを一般に公開することには根本的な問題がある。モデルが人間の感情をシミュレートするコンテンツを生成する場合、これは本質的に操作的（manipulative）である。"



３．プラットフォーム規制と（生成）AI：デジタルサービス法（2022年発効）

- EUのプロバイダ責任を規定してきた電子商取引（2000年）を元に、違法・有害情報に対するプラットフォームの責任・責務や透明性のあり方を全面的にアップデート
- 第III章では、媒介サービス事業者一般やプラットフォーム事業者一般に適用されるコンテンツモデレーション透明性・救済規律の他、EU域内で月間アクティブ利用者4,500万人以上を有する「**超大規模オンラインプラットフォーム（very large online platform、VLOP）**」＋「**超大規模オンライン検索エンジン（very large online search engine、VLOSE）**」事業者に、偽・誤情報を含むシステミックリスクの評価・軽減義務を課す
  - 2023年4月25日に17のVLOPと2のVLOSEが指定、2024年2月全面適用開始
- デジタルサービス法の要点
  - ①コンテンツモデレーション：透明性と救済（生成AI以前からの論点）
  - ②データ保護：プロファイリング規制（生成AI以前からの論点）
  - ③VLOP/VLOSE：偽・誤情報を含むシステミックリスクの評価と軽減（特に生成AI以降の論点）

デジタルサービス法の対象事業者区分と主な規律

仲介サービス（IS）：導管、キャッシング、ホスティングの3種類
免責ルール（2章）、連絡先・代理人・利用規約規制等（3章１節）

ホスティングサービス（HS）：利用者提供情報ホスティング全般
違法コンテンツ通知と措置、理由の説明、透明性レポート等（3章2節）

オンラインプラットフォーム（OP）：利用者提供情報の公衆配布
零細・中小義務除外、内部苦情処理、信頼できる騎手、反復侵害者対応、広告規制・透明性、
レコメンデーション透明性、未成年者保護（3章３節）

超大規模OP（VLOP）：EU域内利用者4,500万人以上のOP
システミックリスクの評価と軽減、危機対応メカニズム、独立監査、当局・
研究者へのデータ提供、コンプライアンス体制整備等（3章5節）
欧州委員会による監督・調査・執行・モニタリング（4章4節)

取引OP：消費者と事業者の間の契約締結を可能とするOP
事業者トレーサビリティ等（3章４節）

超大規模オンライン検索エンジン（VLOSE）：EU域内利用者4,500万人以上の検索エンジン
VLOPに課される3章5節の義務とほぼ同様

①コンテンツモデレーション：透明性と救済

3条(t)：「コンテンツモデレーション」とは、自動的か否かに関わらず、媒介サービス提供者が行う、特にサービス受領者が提供する違法コンテンツ又はその利用規約に適合しない情報の検出、識別、対処を目的とした活動をいい、降格、収益不能化、アクセス不能化、削除など、違法コンテンツ又はその利用規約違反情報の利用可能性、可視性、アクセス性に影響を与える措置、サービス受領者のアカウントの終了又は停止など、サービス受領者の情報提供能力に影響を与える措置を含む。

- 利用規約へのコンテンツモデレーションポリシー明記（IS、14条）
  - 利用者提供情報に関する制限の情報（**アルゴリズムによる意思決定と人間によるレビューを含むコンテンツモデレーションのあらゆる方針・手順・手段・ツール**、内部苦情処理システム手続に関する情報を含む）
  - 制限の実施における表現の自由やメディアの自由・多元性、その他基本権等の利益への配慮義務
  - VLOP/VLOSEは全サービス提供加盟国の言語で当該情報を提供
- 透明性レポート（IS~VLOP段階、15条他）→VLOPの第一次レポート
  - 当局命令・対応、違法・規約違反別の通知・対応件数と対応時間、コンテンツモデレーション担当者訓練内容、**自動処理のエラー率指標とセーフガード措置等**（※VLOPは加盟国の公用語ごとに整理）
- 削除等の理由の説明（IS、17条）　→欧州委による集約データベース
  - コンテンツ削除・降格やアカウント停止等を受けた利用者への明確かつ具体的な理由説明
- 削除等に異議がある場合の内部苦情処理システム整備（OP、20条）
  - 削除やアカウント停止等の判断が誤っていた場合の回復等
- さらに異議がある場合の裁判外紛争処理の利用（OP、21条）
  - 紛争処理機関に対する当局の認定等

②データ保護：プロファイリング規制

- PF上の**ターゲティング広告**のパラメータ等の明示（OP~VLOP段階、26条他）
- **レコメンダーシステム**のパラメータ明示とユーザーによる修正可能性（VLOPはプロファイリングに基づかない選択肢の提供を含む）（OP～VLOP、27条他）
- **GDPR特別カテゴリー個人データのプロファイリング広告利用禁止（OP、26条3項）**
- **青少年保護と未成年個人データのプロファイリング広告利用禁止（OP、28条2項）**

- ※ダークパターンの禁止（OP、25条）：「サービス受領者を欺いたり操作したりするような方法で、又はその他の方法でサービス受領者が自由かつ情報に基づく決定を行う能力を実質的に歪めたり損なったりする方法で、オンライン・インターフェースを設計、組織、運用しないこと」

③VLOP/VLOSE：偽・誤情報を含むシステミックリスクの評価と軽減

- VLOP/VLOSEは、自らのサービスがもたらしうる違法コンテンツ流布、**基本権（特に**人間の尊厳、プライバシー、個人データ保護、表現・情報の自由、非差別、児童の権利、**消費者保護**）、市民言説と選挙、ジェンダー暴力・公衆衛生・青少年保護等への影響**等の「システミックリスク」を自ら特定・分析・評価し（34条）、合理的・比例的・効果的な軽減措置を採る義務（35条）**と、公共の安全・公衆衛生への重大な脅威における危機対応メカニズムにおいて出される欧州委員会の要請決定の対象となる（36条）
- 欧州委員会が奨励・推進・招請して策定する、**行動規範（codes of conduct）（45条）**や危機プロトコル（48条）**を通じて具体化する共同規制メカニズム**
  - デジタルサービス法採択以前から偽情報行動規範が策定、2022年6月の改訂によりディープフェイク等への対応も含まれる
- 34条・35条の義務及び、行動規範・危機対応プロトコルの遵守について、年1回以上の独立監査を受ける義務（37条）
  - 評価・緩和措置検証のための外部研究者データアクセス提供義務（40条）

４．いくつかの論点

- 対応すべき消費者リスクの区分・特定
  - 製品安全（生成AI以前から）
  - プロファイリング（生成AI以前から）
  - 偽・誤情報にとどまらない操作・欺瞞（生成AI以降）
- プラットフォームレイヤーへの着目
  - AIとレコメンダーやプロファイリング、コンテンツモデレーション（生成AI以前から）
  - AI生成コンテンツの流通への対応（生成AI以降）
- 新しい自主・共同規制アプローチ
  - 透明性とモニタリング
  - 「システミックリスク」の特定と緩和