AIとデジタル政策 ver2.0

2023年8月21日

谷脇　康彦

人間中心のAI社会原則(2019年3月）

1. 人間中心の原則
2. 教育・リテラシーの原則
3. プライバシー確保の原則
4. **セキュリティ確保の原則**
5. **公正競争確保の原則**
6. **公平性、説明責任及び透明性の原則**
7. イノベーションの原則

■**特定の企業にAIに関する資源が集中した場合**においても、**その支配的な地位を利用した不当なデータの収集や不公正な競争が行われる社会であってはならない**。

■AIの設計思想の下において、人々がその人種、性別、国籍、年齢、政治的信念、宗教等の多様なバックグラウンドを理由に不当な差別をされることなく、**全ての人々が公平に扱われなければならない**。

■ AIを利用しているという事実、AIに利用されるデータの取得方法や使用方法、AIの動作結果の適切性を担保する仕組みなど、**用途や状況に応じた適切な説明**が得られなければならない。

■AIを効率的かつ安心して社会実装するため、（中略）**倫理的側面**、経済的側面など幅広い学問の確立及び発展が推進されなければならない。

（注）AI開発利用原則については「現在、多くの国、団体、企業等において議論されていることから、我々は早急にオープンな議論を通じて国際的なコンセンサスを醸成し、非規制的で非拘束的な枠組みとして国際的に共有されることが重要である」と記載。

（出典）総合イノベーション戦略推進会議「人間中心のAI社会原則」（2019年3月）

米国におけるAIガバナンスの動き

■2023年7月、大統領府は生成AIに関する「自主的なAIコミットメント」について**AI関連7社※と合意**(拘束力はなく、企業の取り組みの具体策もない)。

※Amazon, Anthropic, Google, Inflection, Meta, Microsoft & OpenAI

安全性

1) AIのモデル・システムの部内・対外的な**Red Teaming** の実施

2) **情報共有**の促進

セキュリティ

3) **サイバーセキュリティ**や内部脅威に対するセーフガードへの**投資**

4) **第三者による脆弱性の発見・報告を促す仕組み**

信頼

5)コンテンツがAI製であることがわかる**電子透かしなどの開発**

6)Aiのもつ能力や限界(公平性やバイアスのような社会的リスクを含む)に関する**情報公開**

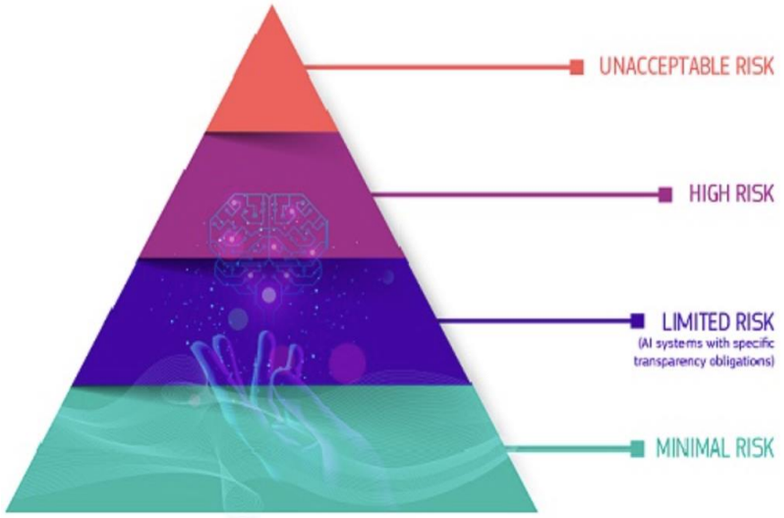
7)有害なバイアスや差別の回避やプライバシー保護のように**Aiがもたらす社会的リスクの研究を優先**

8)気候変動などの社会課題に対応できる**最前線のAIシステム(frontier AI systems)の開発**

■**大統領令の制定**や超党派による**AI法案制定の可能性**の追求などにも言及。

欧州：AI法の制定に向けた動き

リスクベースアプローチ

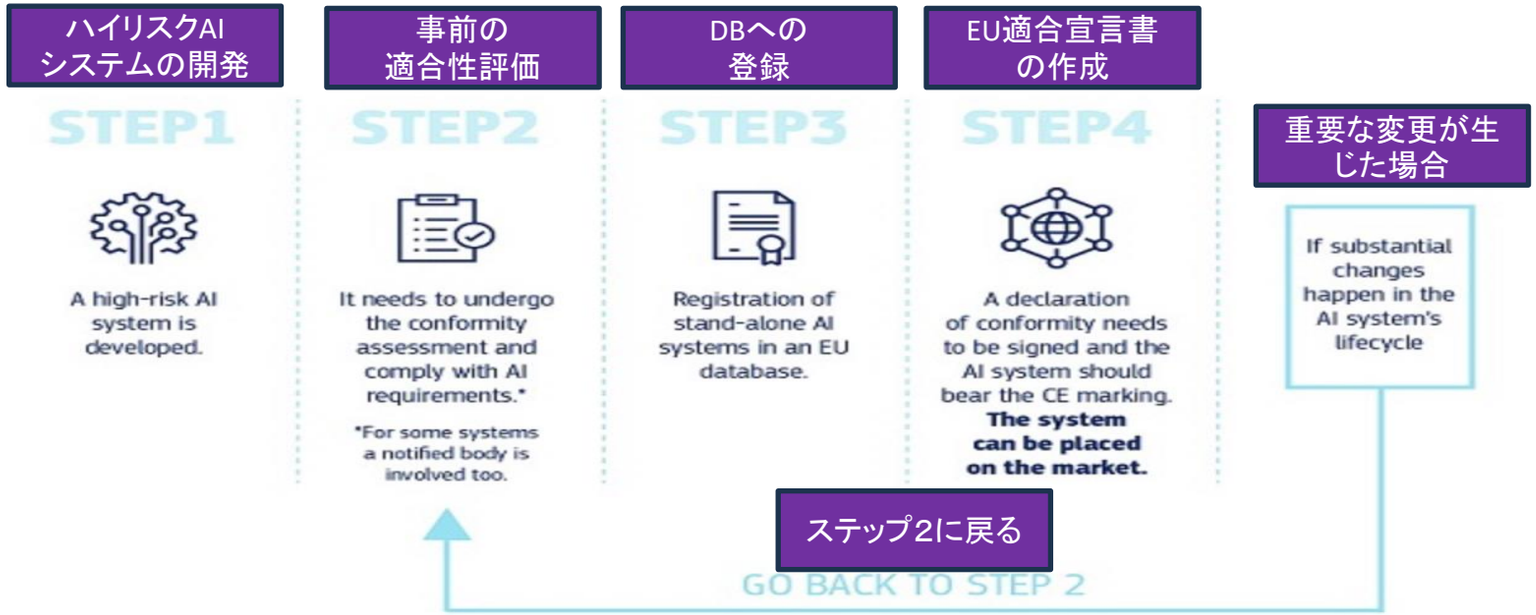


許容できないリスク　禁止

ハイリスク　規制　（規制のプロセスは下図参照）

限定リスク　透明性の義務　AIとのやりとりであることを利用者に知らせることが必要。

最小リスク　規制なし　（EUで利用されている大半のAIはこのカテゴリーとなる旨の注記あり）



（注）GDPRと同様に制裁の域外適用（最大3千万euroか全世界売上高の6%どちらか高い金額）

(出典) EU “Regulatory Framework Proposal on Artificial Intelligence” https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

中国におけるAI規則の制定

■2023年8月、中国は「生成人工知能サービス管理のための規則」を施行。

第4条
1) **法律や行政規則で禁止されているコンテンツの作成禁止**(※)
2) **差別を防**止するための効果的な対**策**
3)知財・企業倫理の尊重ほか、**独占的・不法な競争行為の禁止**
4)他人の肖像権、名誉、プライバシー、個人情報に関する**権利侵害の禁止**

第6条
・生成AIに関する**国際ルールの策定への参加**

第12条
・提供者は、(AIにより)生成されたコンテンツを**識別可能**とする。

第21条
・違反が重大な場合はサービス提供停止を命令。犯罪であれば刑事責任を追及。

(※)社会主義の中核的価値観を遵守し、国家権力の転覆の煽動、社会主義システムの転覆、国家の安全と利益を危険にさらす(中略)など、法律や行政規則で禁止されているコンテンツの作成を禁止＝**【実質的に海外の生成AIを排除】**

(出典)中国政府HPを元に読売新聞記事(23.8.16)などで補足。

人工知能及び自律性の責任ある軍事利用に関する政治宣言

(US DoS, “Political Declaration on Responsible Use of Artificial Intelligence and Autonomy”, Feb 2023)

”軍事領域における責任あるAIに関する会議[REAIMSummit]”(2023年2月＠ハーグ）において、米国務省が提案。軍事分野におけるAIの開発・配備・使用に際し、自主的に遵守し、そのコミットメントをオープンにすることを提案。

- 軍事AI(military AI capabilties)が国際法（特に国際人道法）の義務に合致した形でのみ使用されることを保証するため、法的審査 (legal reviews)などの効果的な措置を実施
- 核兵器の使用に関する重要な情報付与や判断は、人間による管理と関与を維持
- 兵器システムを含むすべての軍事AIの開発・配備は、高官(senior offioials）が監督
- 軍事AIの責任ある設計・開発・配備・使用に関する原則を採択・公表
- 軍事AIの開発・配備・使用は、適切なレベルの職員が判断
- 軍事AIの意図しない偏り(unintended bias)を最小化する対策を実施
- 監査可能(auditable)な方法やデータソース、設計手順、文書によって軍事AIを開発
- 軍事AIを使用する職員、及び使用を承認する職員は、その能力と限界を十分に理解し、その使用について適切な判断を行うことができるよう訓練。
- 軍事AIの安全性・セキュリティ・有効性についは、ライフサイクル全体にわたって厳格なテストと保証の対象とし、自己学習による軍事AIは、重要な安全機能が低下していないことを確認する監視プロセスに従うこと。
- 意図しない結果を検出・回避・解除できるよう設計する等のセーフガードを導入。
- 軍事AIの開発・配備・使用に関する議論を継続し、他の適切なコミットメントを見出すよう努力。

(出典) https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/

AIとデジタル政策

AIガバナンスをどう確立するか。

・「公平性、説明責任及び透明性の原則」に照らし、AIのブラックボックス化を回避する手法の明確化が必要。
・何によって担保するか---ハードローかソフトローか。
・国際的な整合性は確保できるのか。

人口減少下においてAIによる“個別化＆自動化”が必須。区別と差別の切り分けが必要。

・第三者による検証の仕組みをどのように確立するか。

大規模言語モデル（LLM）の多様性をどう確保するか。

・LLMの競争性を確保することでAIガバナンスが有効に機能する可能性。
・LLMのオープン性の確保（APIの開放）が多様な強化学習モデルを生み出す可能性。

AIはサイバー空間における健全性（情報の信頼性）を損なうか。

・Chat GPTの文章の生成は単語と単語の確率的つながりによるため、そこで生じる誤りやAIによるフェイクニュース生成がサイバー空間の健全性を損なわないか。

サイバー攻撃にAIを活用しないとするノーム（規範）は成立するか。

・脆弱性発見やマルウェア開発にAIを使うことで「自動戦争」につながるのではないか。

G7によるAI原則の検討
（2023年5月20日、広島G7首脳コミュニケ）

国や文化を超えてますます顕著になっているAIの機会及び課題について**直ちに評価する必要性**を認識し、人工知能グローバルパートナーシップ（GPAI）が実践的なプロジェクトを実施することを奨励。

生成AIに関する議論のために、包摂的な方法で、OECD及びGPAIと協力しつつ、**G7の作業部会を通じた、広島AIプロセスを年内に創設**する。

これらの議論は、**ガバナンス、著作権を含む知的財産の保護、透明性の促進、偽情報を含む外国からの情報操作への対応、これらの技術の責任ある活用**といったテーマを含み得る。

AIを巡る議論において基本となる問題意識

- 具体的なAIガバナンス論が必要
- 人権保護（差別禁止、プライバシー保護等）や安全保障関連の議論に優先順位
- 規制論は、AIが”moving target”であることに注意
- ハードローありきではなく、ソフトローや技術実装を含めて柔軟に考える必要
- 価値観に関わる問題もあり、広範なステークホルダーを巻き込む議論を