

人工知能関連技術の研究開発及び活用の 適正性確保に関する指針

令和7年 12月 19日
人工知能戦略本部決定

目 次

1 我が国における適正性確保に関する基本的な考え方	2
(1) 本指針の位置付け	2
(2) 本指針における適正性確保の考え方	2
(3) 適正性確保のための基本方針	4
2 研究開発機関及び活用事業者が特に取り組むべき事項	5
(1) A I ガバナンスによる俯瞰的な適正性の確保	5
(2) ステークホルダーとの信頼関係の構築に向けた透明性の確保	5
(3) 十分な安全性の確保	6
(4) 事業継続性確保による安全な環境の維持	6
(5) A I のイノベーションの基盤となるデータの重要性を踏まえたステークホルダーへの配慮 ..	6
3 国及び地方公共団体が特に取り組むべき事項	7
(1) A I の積極的かつ先導的な活用によるイノベーションの促進	7
(2) 社会全体におけるA I リテラシーの向上	7
(3) A I ガバナンスの在り方の検討	7
(4) 行政としてのアカウンタビリティを果たすこと	8
4 国民が特に取り組むべき事項	9
(1) 人間中心の原則に基づくA I の責任ある利用	9
(2) A I リテラシーに基づく適切な利用	9

1 我が国における適正性確保に関する基本的な考え方

(1) 本指針の位置付け

人工知能関連技術の研究開発及び活用の推進に関する法律（令和7年法律第53号。以下「A I法」という。）第13条に基づく本指針は、信頼できるA Iの実現に向けて、事業者、国民等の全ての主体¹におけるA Iの研究開発・活用の適正な実施に係る自主的かつ能動的な取組を促すために、国際的な規範の趣旨に即して策定するものである。

本指針の構成としては、1で全ての主体におけるA Iの研究開発・活用の適正性確保に必要なとなる主な要素と基本方針を示し、2以降で各主体が1を前提として特に取り組むべき事項を記載する。全ての主体は、これに基づき、適正性確保に必要なとなる主な要素を認識、理解することが求められる。また、取り組むべき事項について、各主体の規模や立場、A Iがもたらすリスクに応じて、その時点で適用し得る技術や知見を踏まえて適切な水準で対応することが求められる。

我が国は、信頼できるA Iの開発、活用、普及に向けて、本指針を中心とした枠組みを、国際モデルとなるよう展開するとともに、A Iに係る国際的なルール形成を行う枠組みである「広島A Iプロセス」を牽引してきた実績²を踏まえ、引き続き国際的な議論を主導しながら、A Iガバナンスの構築において国際協調を図る。

(2) 本指針における適正性確保の考え方

A Iは、経済成長や国民生活の発展に寄与するものであることから、その社会実装を進めイノベーションを促進していくことが重要である一方、A Iには様々なリスクがある。例えば、誤判断やハルシネーション³等の技術的リスク、偽・誤情報の生成・拡散、偏見・差別の助長、犯罪への利用、過度な依存、プライバシー・財産権の侵害、環境負荷の増大、雇用・経済不安等の社会的リスク、さらにはサイバー攻撃等の安全保障上のリスクがあり、これらのリスクはA Iの技術進歩とともに変化したり、未知のリスクが発生したりする可能性があり、リスクに対する社会的な受容水準も変化し得る。

¹ A I法第4条から第8条において責務が規定される、国、地方公共団体、研究開発機関、活用事業者、国民を指す。

² 2023年5月のG7広島サミットにおいて「広島A Iプロセス」を立ち上げ、G7日本議長国下の成果物として、高度なA Iシステム開発・利用に関する「広島A Iプロセス包括的政策枠組み」をとりまとめた。2024年5月には、広島A Iプロセスの精神に賛同する国々・地域の自発的な枠組みである「広島A Iプロセス・フレンズグループ」が立ち上げられ、60の国・地域が参加している（2025年12月時点）。また、2025年2月には、「国際行動規範」の遵守状況をA I開発者自らが自主的に報告、公表する「報告枠組み」の正式運用が開始されており、24組織が回答を提出している（2025年12月時点）。

³ 生成A Iにより、事実とは異なることがもってまわしく出力されることをいう。

このため、本指針では、適正性確保に当たって、適正性についての一義的な定義や絶対的な水準を定めるものではなく、各主体が研究開発、活用するAIの特性、用途、目的や、自身の立場、社会的役割等を踏まえて自主的に取組を進めるという考え方の下、「人間中心のAI社会原則」（平成31年3月29日統合イノベーション戦略推進会議決定）に掲げられた理念を踏まえ、その際に考慮すべき主要な要素を以下のとおり示す。

- 人間中心
 - ◇ 人間の尊厳や基本的人権を尊重すること。また、法令を遵守すること。
 - ◇ 誰もがAIの恩恵を享受できるよう、多様性、包摂性を尊重し、多様な人々の幸福の追求による包摂的な成長を目指すこと。
 - ◇ AIを活用する範囲や条件については、人間自らが最終的な判断を行うこと。
- 公平性
 - ◇ AIの活用によって、社会に不当な偏見や差別を生じさせたり、助長したりしないこと⁴。
- 安全性
 - ◇ AIの活用によって、生命、身体、財産等⁵に危害を及ぼさないようにすること。
- 透明性
 - ◇ AIに対する信頼性が向上するよう、必要かつ技術的に可能な範囲での情報の開示、事後的な検証可能性の確保等により、透明性を適切に確保すること⁶。
- アカウンタビリティ⁷
 - ◇ AIがもたらす社会的影響を踏まえ、責任の所在の明確化、責任を果たすための仕組みの構築等により、技術的、制度的、社会的観点からアカウンタビリティを合理的な範囲で果たすこと。
- セキュリティ
 - ◇ 不正な操作によるAIの意図しない動作や停止をはじめとするAIのセキュリティ上のリスクを低減させるよう、AIのセキュリティを適切に確保すること。
- プライバシー・個人情報
 - ◇ 取り扱うデータの重要性等に応じてプライバシーを尊重し、適切に保護すること。また、個人情報保護法等関連法令を遵守すること。

⁴ AIの活用によって生じ得るバイアス、ジェンダーギャップ、情報操作等により公平性を損なわないようにすることを含む。

⁵ ディープフェイク技術によるフェイク動画、性的加工画像、他人になりすました音声等を用いた脅迫や名誉棄損により危害が及ぼされ得る自由、名誉を含む。

⁶ AIの挙動、入力から出力を生成するプロセスの解明等を進め、AIの出力に寄与するアルゴリズムへの理解を深めることも重要である。

⁷ 個人や組織が自らの行動や決定に対する責任を持ち、その責任を果たすための行動をとることを意味する。

- 公正競争
 - ◇ 特定の者にA Iに関する資源が集中した場合においても、その有利な立場を利用した不当なデータの収集を含む不公正な取引が行われないようにするなど、公正な競争の促進に貢献すること。
- A Iリテラシー
 - ◇ A Iがもたらすリスクの社会的受容可能な水準は変わり得ることを認識し、便益の最大化とリスクの抑制を図れるよう、知識・能力を身に付けるとともに、倫理観を保持すること。
- イノベーション
 - ◇ 環境負荷の低減を含む持続可能性を確保しつつ、イノベーションの促進に貢献するよう努めること。
 - ◇ 社会課題の解決に資するA Iの技術開発に取り組むこと。
 - ◇ A Iの活用を阻害する要因の改善を図ること。

(3) 適正性確保のための基本方針

(2)で示した考え方を踏まえ、適正性を確保するために取り組むべき基本方針を以下のとおり示す。

① リスクベースでのアプローチ

A Iがもたらすリスクを特定・評価し、A Iを利用する分野や目的を踏まえた影響度合いに応じて、適切な対策を講じる⁸。

② ステークホルダーの積極的な関与

A Iがもたらす便益、リスク等により影響を受ける主体（以下「ステークホルダー」という。）⁹がA Iガバナンスに積極的に関与し、他のステークホルダーと協働して課題解決に取り組む。

③ 一貫通貫でのA Iガバナンスの構築

A Iがもたらす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、便益を最大化するため、研究開発から社会実装までが近接するA Iの各段階を一体的に捉えたA Iガバナンスを構築する。

④ アジャイルな対応

A Iの技術進歩の速さや、予見可能性、説明可能性が十分でないことを踏まえ、変動し得るリスクに対してP D C A（計画・実行・評価・改善）サイクルを回しながら、柔軟・迅速（以下「アジャイル」という。）に対応し、A Iガバナンスの成熟度を高めていく。

⁸ レッドチーム等の様々な手法を組み合わせ、多様な内部テスト手段や独立した外部テスト手段を採用することや、特定されたリスクや脆弱性に対処するための適切な措置を実施することが望ましい。

⁹ 例えば、A Iのイノベーションの基盤となるデータの保有者、A Iの生成物を取り扱う者等、A Iの活用によって影響を受けるが、直接は関与していない関係者も含まれ得る。

2 研究開発機関及び活用事業者が特に取り組むべき事項

AIを活用した製品、サービスの開発、提供をする活用事業者¹⁰は、その開発、提供したAIが多くの主体に影響を及ぼし得ることを踏まえ、国際的な規範¹¹、国際規格¹²、各種ガイドライン等¹³を活用しつつ、1(2)に示す適正性確保に必要となる主な要素に関して、特に以下の事項に取り組む。

また、AIの研究開発機関¹⁴は、その開発したAIを第三者に提供する際は、1(2)に示す適正性確保に必要となる主な要素に関して、特に以下の事項に取り組む。

(1) AIガバナンスによる俯瞰的な適正性の確保

AIの設計・開発・提供・実装等のライフサイクル全体で、リスクの特定・評価・対処をするための組織的なプロセス（経営層の関与したモニタリングや評価の仕組み、情報の適切な開示、教育・研修の実施等）を含むAIガバナンスを構築¹⁵・運用・継続改善し¹⁶、AIがもたらす便益を最大化しつつ、そのリスクを受容可能な水準で管理する。

(2) ステークホルダーとの信頼関係の構築に向けた透明性の確保

学習データの出所と出力される生成物について、知的財産、プライバシー等の保護の適切な実施を含め、ステークホルダーとの信頼関係を構築するためにも、合理的な範囲で説明可能性を確保する¹⁷。

また、AIを提供する際、AIの適正な利用を可能にするための情報（AIの仕組み・限界、禁止事項、学習するデータの収集ポリシー、出力の信頼性に関する注意喚

¹⁰ AI法第7条に規定する活用事業者をいう。海外事業者も含む。

¹¹ 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（原文：Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems）、全てのAI関係者向けの広島プロセス国際指針（原文：Hiroshima AI Process International Guiding Principles for All AI Actors）等

¹² AIマネジメントシステム（ISO/IEC 42001）等

¹³ 内閣府ウェブサイト参照（国内外のAIに関する規範、ガイドライン等を示す予定。）

¹⁴ AI法第6条に規定する研究開発機関をいう。研究開発機関のうち大学については、AI法第6条第2項に定めるとおり、研究者の自主性の尊重その他の大学における研究の特性に配慮するものとする。

¹⁵ ゼロベースからAIガバナンスを構築することに限らず、既存のITシステム等に適用されているガバナンスプロセスを活用することも有用である。

¹⁶ この際、AIに関するリスクの回避策や、リスクが顕在化した場合の対応を含む信頼性の高い組織を構築するため、「広島AIプロセス」報告枠組みの活用や国際規格（AIマネジメントシステム（ISO/IEC 42001）等）に基づくマネジメント体制の整備・運用をすることにより、AIに関連するリスクの管理・制御を図り、社会的・倫理的責任を果たすことが可能と考えられる。これらの取組を積極的に開示・説明することが、企業価値の向上や競争優位性の確保にもつながるものと期待される。

¹⁷ 学習データ等の適切な透明性を確保するため、AIの出力の際に根拠とした情報（ウェブサイト等）を表示する。また、学習データ等の開示が求められた際は、可能な限り対応することが望ましい。技術的な制約によりAIの出力と学習データの関係性を特定することが困難な場合や、開示が求められた学習データ等が営業秘密に該当する場合などにおいても、まずは真摯に検討、協議することが期待される。

起等¹⁸⁾をそのA Iの利用者に提供する。

(3) 十分な安全性の確保

A Iを悪用したサイバー攻撃や詐欺をはじめとする各種犯罪その他の違法行為が行われるリスクを特定・評価し、適切な対策を講じる。

また、ハルシネーションや偏見・差別の助長、偽・誤情報等（ディープフェイク技術によるフェイク動画、性的加工画像等）の拡散等につながるA Iによる不適切な出力の抑制、A Iの意図しない動作や誤作動の防止をするため、最新の技術と知見を駆使して、解決、改善に向けて取り組む。特に、A Iで生成された偽・誤情報等の拡散が深刻なリスクとなっていることを踏まえ、A Iの生成物であることが判断できる技術（電子透かし、来歴管理、API¹⁹⁾等）の開発に努め、必要に応じて実装する。

(4) 事業継続性確保による安全な環境の維持

A Iを用いたシステムの運用者やサービスの提供者は、これらに障害が生じた場合に備え、損害を最小限にとどめ、早期復旧するために、平常時に行うべき活動や緊急時の事業継続のための方法、手段等を定めた事業継続計画をあらかじめ策定する。

(5) A Iのイノベーションの基盤となるデータの重要性を踏まえたステークホルダーへの配慮

A Iのイノベーションには、質の高いデータを確保し、それらを適正に活用することが重要である。これを踏まえ、質の高いデータが充実し、信頼できるA Iが開発、提供されることにより、新たな創作活動等が促進されるという好循環を実現するため、A Iを開発、提供する事業者は、データの利用状況に応じて、知的財産等のデータ保有者等のステークホルダーと、データの適正な活用の在り方等について継続的なコミュニケーションを図る。また、特に、社会的影響力の大きいA Iを開発、提供する事業者は、知的財産等のデータ保有者等に対する利益還元のエコシステムや安心して創作活動等ができる環境の構築に向けた方策の検討、実施に努める。

¹⁸⁾ 雇用、人事評価等における不当な偏見・差別や、偽・誤情報の拡散等につながる利用者の不適切な行為を防止するための注意喚起、利用者からの問い合わせに対応する窓口、連絡先等を含む。

¹⁹⁾ Application Programming Interface。異なるアプリケーション（ソフトウェア）やシステムを連携させ、通信やデータ交換を行うための仕組み。

3 国及び地方公共団体が特に取り組むべき事項

国は、1(2)に示す適正性確保に必要となる主な要素に関して、特に以下の事項に取り組む。

地方公共団体は、1(2)に示す適正性確保に必要となる主な要素に関して、置かれた環境や課題が多様であることを踏まえ、地域の実情に応じて、特に以下の事項に配慮しつつ、必要な対応を行う。

なお、AIを開発、提供する際は、2に示す事項についても取り組む。

(1) AIの積極的かつ先導的な活用によるイノベーションの促進

国又は地方公共団体におけるAIの活用事例や留意点を広く周知することがAIの普及促進にも効果的であることを踏まえ、自ら積極的かつ先導的にAIを活用する。また、公共調達における開発実証機会の提供も進める。

(2) 社会全体におけるAIリテラシーの向上

国及び地方公共団体は、国及び地方公共団体の職員はもちろんのこと、全ての主体が、倫理、法令、人権、安全等に関する課題を理解し、責任ある利用者としての自覚をもって行動できるように、社会全体におけるAIリテラシーの向上を図ることが求められる。

このため、常にAIの最新の技術動向や活用実態を把握し、リスク及びその対応策を検討して、ステークホルダーの自主的な取組を促すための考え方を提示する。また、事業者、国民等におけるAIの研究開発・活用における適正性確保に向けて、生成AIの基本的な使い方や注意点を学べるコンテンツの提供、社会人向けの生成AIスキル・知見の習得支援等、教育・ガイダンスを積極的に推進する。

(3) AIガバナンスの在り方の検討

国内外におけるAIガバナンスを巡る動向を注視し、AIガバナンスの在り方を継続的に検討し、対応する。本指針及び各種ガイドライン等も、AIの技術進歩による社会の変化をとらえ、アジャイルかつ継続的に見直す。この際、各種ガイドライン等は、本指針の趣旨と整合するとともに、事業者、国民等にとってわかりやすいものとなるよう、適宜、点検・見直しを行う。

また、様々な局面におけるAI導入の障壁を低減するため、AIを活用する際に想定・発生し得る課題に対して、その責任の所在等に関する解釈適用上の論点及び考え方を整理するとともに、判例等を踏まえ可能な限り解釈を明確化するよう努める。

さらに、A Iは国境を越えて展開されるため、国内だけでなく、国際的なガバナンスが不可欠であり、相互運用性の確保にも配慮しつつ、A Iガバナンスの構築を主導する。

(4) 行政としてのアカウントビリティを果たすこと

行政においてA Iを活用する際、行政の信頼性を確保するため、求められる水準を十分に考慮した適切なリスク対策等を実施し、可能な限り判断の根拠等が不明瞭にならないよう国民へのアカウントビリティを果たす。

また、各府省庁において、A Iガバナンスの責任者を任命する²⁰。地方公共団体においては、A Iの適正な活用、リスク管理における責任者を明確化する。

²⁰ 「行政の進化と革新のための生成A Iの調達・利活用に係るガイドライン」（令和7年5月27日デジタル社会推進会議幹事会決定）においては、各府省庁における生成A I利活用方針を策定・推進し、組織全体の利活用状況とリスク管理等を統括管理する者としてA I統括責任者（CAIO）を設置することとされている。生成A I以外のA Iを活用する際も、必要に応じて責任者を明確にすることが望ましい。

4 国民が特に取り組むべき事項

国民は、1(2)に示す適正性確保に必要となる主な要素に関して、以下に特に配慮した対応を行う。

(1) 人間中心の原則に基づくA Iの責任ある利用

国民はA Iを利用する主体として、A Iの利用により法令への抵触、加害行為をす
るおそれがあることを踏まえ、法令を遵守する。

また、A Iの利便性のみならず、倫理、法令、人権、安全等に関する課題を理解し、
責任ある利用者としての自覚をもって行動するよう努める。

(2) A Iリテラシーに基づく適切な利用

A Iの特性や仕組みを正しく理解し、能動的にA Iリテラシーを身に付けるよう努
める。

その上で、A Iを利用する際は、得られる情報の出所、正確性等を理解し、人間の
判断、責任の下で意思決定を行うとともに、不当な偏見・差別、誹謗中傷、偽・誤情
報の拡散等を目的とした不適切な行為を行わない。また、A Iの生成物（文章、画像、
音声、動画等）は、社会的、法的に適切な形で利用する。