AIガバナンス行動目標（案）

一般社団法人AIガバナンス協会
2024年2月 策定
2025年11月 改定

前文

生成AIをはじめとする技術的なブレークスルーが進む今日、AIの社会実装は急速に進み、企業、政府、市民社会等のさまざまなステークホルダーが不可避的にAIと関わりながら活動する時代が到来しつつあります。今や、社会課題の解決や持続可能なビジネスを進める上で、AIは必須の存在です。

しかし、技術・社会の変化の中で、AIのもたらすリスクも広く認識されるようになりました。特に生成AIと関連して注目されるようになったリスクとして、誤情報・偽情報の生成、AI技術の悪用、出力へのバイアスの反映、AIへの心理的依存といった課題があります。また、中長期的にはAIの活用が前提となる社会システムが形作られる中、現在はまだ明らかになっていないシステミック・リスクが発生する可能性もあります。

こうした背景のもと、AIの活用を目指す主体に対して、AIのリスクを管理し、その便益を最大化するための取組――「AIガバナンス」を求める社会的要請も高まっています。

私たち一般社団法人AIガバナンス協会（AIGA）は、政府の掲げる「人間中心のAI社会原則」において示される「人間の尊厳が尊重される社会」「多様な背景を持つ人々が多様な幸せを追求できる社会」「持続性ある社会」という基本理念を尊重し、AIガバナンスの普及・推進という手段によってそれらの社会像の実現を目指します。その際、特に以下の3つの価値を重視します。

- **社会的な価値の実現**: AI活用は、人間の尊厳、多様性、安全性及び人間を取りまく環境の保護といった社会的な諸価値を尊重して進められるべきであること
- **マルチステークホルダーでの信頼構築**: AIシステムを取り巻く全ての関係者が、各主体の中での議論に留まらないステークホルダーとの開かれた対話を通じて、相互の信頼関係を築き、維持するべきであること
- **イノベーションの促進**: AIガバナンスは、AIの活用を単に抑制するのではなく、イノベーションを促進し、新しいサービスの創出、社会課題の解決、生産性の向上を実現することを目的として実装されるべきであること

AIGAは日本社会において適正なAI活用を推進する団体として、以上の価値を実現するためにこの「AIガバナンス行動目標」を策定し、本行動目標に従ってAIガバナンスの社会実装を進めていくことを宣言します。

これらの行動目標はいずれもAIGA及びAIGA会員が今後実現していくべきビジョンを示したものであり、現在の実装状況を問わず、同じ目標を共有する主体が幅広くコミットメントを示すことを見据えて策定するものです。

AIGAは、以下5点の「AIGA基本方針」を策定します。AIGAは、その活動全体を通じて、以下に掲げる行動目標の実現を目指します。

1. ***AIガバナンスの民主化***: AIGAは、前文に掲げた価値に沿ったAIの活用やガバナンスのあり方を議論する上で、AI開発者・提供者・利用者といったプレイヤー、政府、学界、それを取り巻く市民社会等の多様なアクターが議論に参画することの重要性を理解し、マルチステークホルダー間での真摯な情報共有や、開かれた議論の実現に向けて取り組みます。その際特に、AI活用の現場の多様な声を公的な場に届け、あるべきAIガバナンス像の社会実装を見据えた「地に足のついた」議論を目指します。

2. ***横断的な共通認識の醸成***: AIGAは、各組織内に閉じたAIガバナンス構築だけでは対外的なトラスト確保が困難な場合があり、主体や業界の垣根を超えた共通認識の醸成が重要であることを認識し、外部ステークホルダーや第三者によるリスク検証も交えて、リスク認識の共有・アップデート、AIガバナンスに関する客観的な標準・評価基準の検討や事例の共有、及び人材育成のあるべき姿の検討を進めます。また、バリューチェーン上のプレイヤー間で連携しながら、今後新たに発生するリスクも見据え、未来志向での技術開発やリスク対策の検討を行います。

3. ***アジャイル・ガバナンスの実装***: AIGAは、AIに関する技術・社会の動向が絶えず変化すること、また、AIという技術そのものが外部環境やデータの変化に大きな影響を受けることを踏まえ、不断にAIガバナンスのあり方を見直しながら改善を図る「アジャイル・ガバナンス」の考え方を重視します。各主体のAIガバナンス方針や取組内容が目指す理念に照らして適切なものになっているかを継続的に評価・検証し、必要な範囲で目標・手段の双方に見直しを加えながら、AIガバナンスの社会実装を後押しし、イノベーションによる積極的な価値創造と、発生しうるリスクへの対策の両立を目指します。。

4. ***リスクベースアプローチ***: AIGAは、イノベーションを阻害せず社会的な価値を実現するための方策としてのリスクベースアプローチを重視します。必ずしも一律のアプローチがあらゆるAIモデル・サービスに適合するわけではないことを認識し、リスク・レベルや業種ごとに異なる多様なガバナンスのあり方を検討することで、さまざまなユースケースにおける健全なAI活用の発展を後押しします。

5. ***国際的議論への積極的参画と、それによる社会規範創出への貢献***: AIGAは、AIガバナンスの実装がグローバルな課題であることや、各国及び国際社会ですでに進められている議論を踏まえ、海外の機関・団体との積極的な情報交換や連携、国際的な議論の場への参画を通じて、AIガバナンスのあるべき像および日本のガバナンスモデルを世界に発信し、AIをめぐる規範創出に対して積極的な役割を果たします。

AIGA会員は、以下の事項について研究・議論や実践に努めます。また、そうした情報を必要な範囲でAIGAを中心とするコミュニティと共有すると共に、「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」等に示された考え方に従って積極的な情報開示や公的な枠組みでのベストプラクティス共有を行い、標準化や法制度の解釈・整備をめぐる議論にも協力することで、社会としての問題解決に貢献します。

6. ***個人情報の適正な取扱い、プライバシーの保護***: AIGA会員は、AIモデル・サービスの開発・提供・利用において、モデルへの入力情報の管理をはじめとする重要情報の保護のための取組を実施し、個人情報保護法等の関係法令を遵守するとともに、データに基づく不当なデータ主体の取扱いの防止、個人情報やプライバシーの保護に努めます。

7. ***知的財産権の保護***: AIGA会員は、AIモデル・サービスの開発・提供・利用において、権利侵害のリスクの低い学習データ・モデルの活用や、モデルへの入力情報の管理等の対策を実施することで、著作権法をはじめとする関係法令の遵守、知的財産の保護に努めます。

8. ***安全性・性能の確保***: AIGA会員は、AIモデル・サービスの出力に技術的に誤りが含まれうることや、AIの品質が運用後も実社会やデータの変化を反映して変化していくこと、またユースケースによっては暴力的・性的な情報をはじめとする有害情報の出力もリスクとなることを踏まえ、適時の性能・リスク評価を実施することで、AIの出力による生命・心身・財産等への危害の防止に努めます。

9. ***公平性の確保***: AIGA会員は、AIモデル・サービスの出力に特定の個人ないし集団への不当な差別が反映される可能性があることを認識し、適用例ごとに追求すべき公平性の要件を熟慮した上で、適時にバイアスの評価等を実施することで、AIの出力による公平性の毀損の防止に努めます。

10. ***悪意ある主体への対策***: AIGA会員は、AIモデル・サービスの開発・提供・利用において、悪意あるユーザによる偽情報生成や詐欺をはじめとする望ましくない活用や、AIをターゲットにした攻撃のリスクが存在すること、そしてそうした攻撃手法が進化・多様化していることを認識し、物理的な対策、データインプット対策、モデルの改善や適切なルール設計といった取組を実施することで、AIのセキュリティ確保と悪用の防止に努めます。

11. **人間-AI間の相互作用をめぐる安全な設計**: AIGA会員は、ユーザがAIと相互作用をする中で発生するAIへの心理的依存や能力開発への悪影響といった負の影響の存在を認識し、過剰な利用への対策、AIリテラシーに課題のあるユーザへの配慮、適切な人間とAIとの間の役割分担の設計等を進めることで、健全なAI活用の環境整備に努めます。

12. ***自律性の高いAIシステムと人とのより良い協働***: AIGA会員は、エージェント型AIをはじめとする高い自律性をもってデータベースや外部システムと接続されるAIシステムの導入が、以上で触れてきたリスクを増幅・複雑化しうることを認識し、AIのデータアクセスや権限の範囲設計、行動履歴の記録、モニタリング等の管理方法を必要に応じて最適化することで、そのメリットを最大化することに努めます。また、幅広いタスクをAIが担う状況を想定し、責任分界の整理等を進めることで、人間とAIの間の適切な協働のあり方を検討します。

13. ***多様化するサプライチェーンリスクの対策***: AIGA会員は、AIモデル・サービスの開発・提供・利用において、国内外の多様なサードパーティベンダー等が関わることを念頭に、データの流通経路やAIモデルの学習状況等のリスク要因について必要な情報を収集し、技術的措置や契約等も含めたリスク対策を推進します。

14. ***継続的なリスク管理と技術を用いたガバナンスの高度化***: AIGA会員は、AIのライフサイクル全体において新たなリスクの発生、リスク耐性の変化、予期せぬ利用方法によるリスクの顕在化といった事態が発生しうることを認識し、AIモデル・サービスの開発・提供において、開発時から運用時に至るまで継続的に、6-13で述べたようなAIリスクの特定、評価、軽減のための適切な措置をとるよう努めます。またその際、人間の目によるガバナンスに限界があることを踏まえ、必要に応じてAIを含む技術を活用したリスク評価やモニタリングの高度化・効率化等を検討します。

15. ***客観的な視点の導入***: AIGA会員は、各企業内に閉じたAIガバナンス構築だけでは十分なリスク検証や対外的なトラスト確保が困難な場合があることを認識し、特にハイリスクなAIモデル・サービスについて、①組織やルールに対する外部の視点からの検証、②第三者による敵対的な役割からのリスク検証（レッド・チーミング）等の技術的な対応、③利用者をはじめとするステークホルダーからのフィードバックを踏まえた改善、といった事項に必要に応じて取り組み、プロセスにおける客観的な視点の導入に努めます。

16. ***透明性とアカウンタビリティの確保***: AIGA会員は、AIモデル・サービスの開発・提供・利用において、AI活用に関する自社の方針を整理し、①AI活用の事実やAIの能力・限界、適切・不適切な利用方法といった事項の開示（透明性）、②各種法令等への準拠状況の可視化や、経営レベルを含む責任体制の明確化（アカウンタビリティ）、③AIの開発・運用実績やリスク検証の結果の文書化（監査可能性）、④活用するデータのトレーサビリティの確保、といった事項に必要に応じて取り組み、AI活用の状況をステークホルダーや社会に向けて説明できる体制の整備に努めます。その際には、自社の事業にとってのリスクのみならず、外部ステークホルダーや地球環境への影響についても考慮します。

17. ***教育・リテラシー向上の推進***: AIGA会員は、AIモデル・サービスの開発・提供・利用の各段階に幅広い人材が関与することを認識し、AIの開発・活用の方法、技術的な仕組みや特性、事業にもたらすリスクと対策等について、必要なリテラシーの度合いに応じた教育・人材育成に努めます。また、ユーザを含む社会に対しても、AIリスクやその対策についての情報開示や啓発を進めます。