参考資料8

AI 事業者ガイドライン

（第 1.1 版）

令和 7 年 3 月 28 日

総務省　　経済産業省

目次

はじめに

AI 関連技術は日々発展をみせ、AI の利用機会及び様々な可能性は拡大の一途をたどり、産業におけるイノベーション創出及び社会課題の解決に向けても活用されている。また近年台頭してきた対話型の生成 AI によって「AI の民主化」が起こり、多くの人が「対話」によって AI を様々な用途へ容易に活用できるようになった。これにより、企業では、ビジネスプロセスに AI を組み込むだけではなく、AI が創出する価値を踏まえてビジネスモデル自体を再構築することにも取り組んでいる。また、個人においても自らの知識を AI に反映させ、自身の生産性を拡大させる取組が加速している。我が国では、従来から Society 5.0 として、サイバー空間とフィジカル空間を高度に融合させたシステム（CPS：サイバー・フィジカルシステム）による経済発展と社会的課題の解決を両立する人間中心の社会というコンセプトを掲げてきた。このコンセプトを実現するにあたり、AI が社会に受け入れられ適正に利用されるため、2019 年 3 月に「人間中心の AI 社会原則」が策定された。一方で、AI 技術の利用範囲及び利用者の拡大に伴い、リスクも増大している。特に生成 AI に関して、知的財産権の侵害、偽情報・誤情報の生成・発信等、これまでの AI ではなかったような新たな社会的リスクが生じており、AI がもたらす社会的 リスクの多様化・増大が進んでいる。

そのような背景の中、本ガイドラインは、AI の安全安心な活用が促進されるよう、我が国における AI ガバナンスの統一的な指針を示す。これにより、様々な事業活動において AI を活用する者が、国際的な動向及びステークホルダーの懸念を踏まえた AI のリスクを正しく認識し、必要となる対策を AI のライフサイクル全体で自主的に実行できるように後押しし、互いに関係者と連携しながら「共通の指針」と各主体に重要となる事項及び AI ガバナンスを実践することを通して、イノベーションの促進とライフサイクルにわたるリスクの緩和を両立する枠組みを積極的に共創していくことを目指す。

我が国は 2016 年 4 月の G7 香川・高松情報通信大臣会合における AI 開発原則に向けた提案を先駆けとし、G7・G20、OECD 等の国際機関での議論をリードし、多くの貢献をしてきた。一方、AI に関する原則の具体的な実践を進めていくにあたっては、

- 少子高齢化に伴う労働力の低下等の社会課題の解決手段として、AI の活用が期待されていること
- 法律の整備・施行が AI の技術発展及びその社会実装のスピード・複雑さとの間でタイムラグが発生すること
- 細かな行為義務を規定するルールベースの規制を行うと、イノベーションを阻害する可能性があること

等が指摘されてきた。これらを踏まえ、AI がもたらす社会的リスクの低減を図るとともに、AI のイノベーション及び活用を促進していくために、関係者による自主的な取組を促し、非拘束的なソフトローによって目的達成に導くゴールベースの考え方で、ガイドラインを作成することとした。

このような認識のもと、これまでに総務省主導で「国際的な議論のための AI 開発ガイドライン案」、「AI 利活用ガイドライン～AI 利活用のためのプラクティカルリファレンス～」及び経済産業省主導で「AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1」を策定・公表してきた。そして、このたび 3 つのガイドラインを統合・見直しして、この数年でさらに発展した AI 技術の特徴及び国内外における AI の社会実装に係る議論を反映し、事業者が AI の社会実装及びガバナンスを共に実践するためのガイドライン（非拘束的なソフトロー）として新たに策

定した（「図 1. 本ガイドラインの位置づけ」参照）。従来のガイドラインに代わり、本ガイドラインを参照することで、AI を活用する事業者（政府・自治体等の公的機関を含む）が安全安心な AI の活用のための望ましい行動につながる指針（Guiding Principles）を確認できるものとしている。また、本ガイドラインは、政府が単独で主導するのではなく、教育・研究機関、一般消費者を含む市民社会、民間企業等で構成されるマルチステークホルダーで検討を重ねることで、実効性・正当性を重視したものとして策定されている。

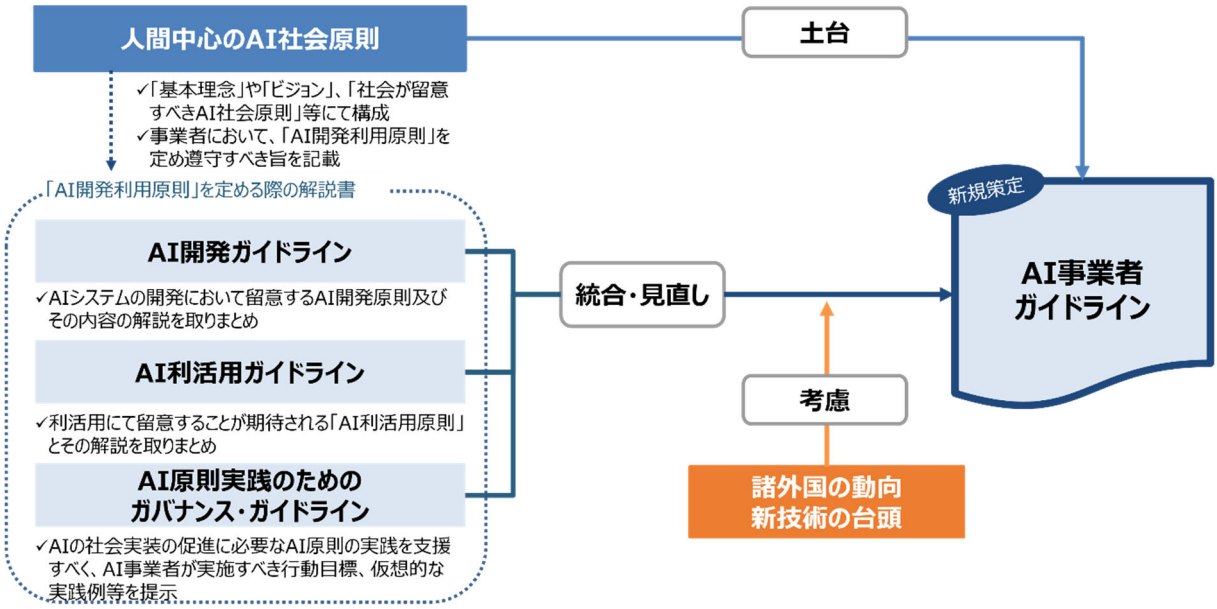

図 1. 本ガイドラインの位置づけ

AI の利用は、その分野とその利用形態によっては、社会に対して大きなリスクを生じさせ、そのリスクに伴う社会的な軋轢により、AI の利活用自体が阻害される可能性がある。一方で、過度な対策を講じることは、同様に AI 活用自体又は AI 活用によって得られる便益を阻害してしまう可能性がある。このような中、予め事前に当該利用分野における利用形態に伴って生じうるリスクの大きさ（危害の大きさ及びその蓋然性）を把握したうえで、その対策の程度をリスクの大きさに対応させる「リスクベースアプローチ」が重要となる。本ガイドラインでは、この「リスクベースアプローチ」にもとづく企業における対策の方向を記載している。なお、この「リスクベースアプローチ」の考え方は、AI 先進国間で広く共有されているものである。

また、AI をめぐる動向が目まぐるしく変化する中、国際的な議論等も踏まえ、本ガイドラインに関しては、AI ガバナンスの継続的な改善に向け、アジャイル・ガバナンスの思想を参考にしながら、マルチステークホルダーの関与の下で、Living Document として適宜更新を行うことを予定している[1]。その中で、社会における AI の成熟度に応じたリスク及びリスクへの対策となる指針並びに実践の内容の更新を検討する（「図 2. 本ガイドラインの基本的な考え方」参照）。

[1] 総務省の AI ネットワーク社会推進会議と経済産業省の AI 事業者ガイドライン検討会にて取りまとめを行う。検討体制は、今後の状況に合わせて適宜見直しを行う。
・AI ネットワーク社会推進会議　https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/index.html
・AI 事業者ガイドライン検討会　https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/index.html

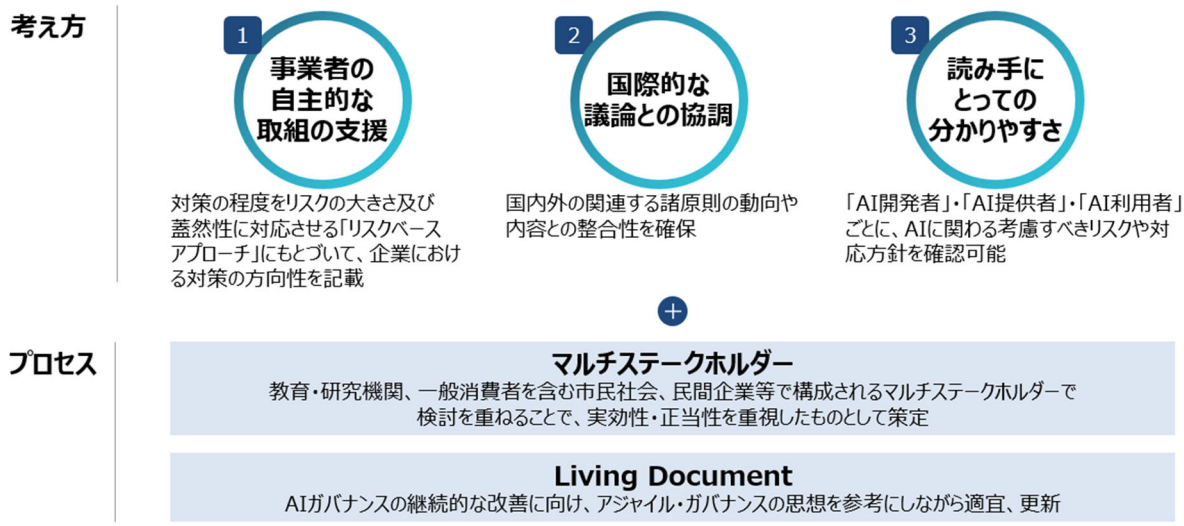

図 2. 本ガイドラインの基本的な考え方

本ガイドラインは、AI 開発・提供・利用にあたって必要な取組についての基本的な考え方を示すものである。よって、実際の AI 開発・提供・利用において、本ガイドラインを参考の一つとしながら、AI 活用に取り組む全ての事業者が自主的に具体的な取組を推進することが重要となる。同時に、AI 活用に取り組む全ての事業者は、AI が社会にもたらす影響の大きさを認識し、人間社会をよりよいものへと発展させるために活用することを意識すべきである。当該取組に対して、社会から不適切又は不十分と評価される場合は、自らの事業活動における機会損失が生じ、事業価値の維持が困難となる事態を招く恐れがあることに留意することが重要となる。このような点に留意することにより、AI による便益の最大化、競争力の強化、事業価値の維持・向上等が可能となる。なお、AI に関係する事業者以外の者、例えば、教育・研究機関に属する者及び一般消費者（未成年を含む）にとっても、AI の活用にあたって参考となる情報及びリスクに関する情報が盛り込まれているため、有用といえる。

本ガイドラインは、様々な事業活動において AI の開発・提供・利用を担う全ての者（政府・自治体等の公的機関を含む）を対象としている。他方で、事業活動以外で AI を利用する者又は AI を直接事業で利用せずに AI システム・サービスの便益を享受する、場合によっては損失を被る者（以下、あわせて「業務外利用者」という）については、本ガイドラインの対象には含まない[2]。ただし、事業活動において AI の開発・提供・利用を担う者から業務外利用者への必要な対応は記載する。また、AI 活用に伴い学習及び利用に用いるデータが不可欠となる。それらのデータを提供する特定の法人及び個人（以下、「データ提供者」という）も同様に本ガイドラインの対象には含まない。データ収集は色々な方法が考えられる中で、本ガイドラインではデータの提供を受ける者・データを入手する者にあたる AI の開発・提供・利用を担う者がデータを取り扱う際の責任を負う形で記載する。以上より、本ガイドラインの対象者は、AI の事業活動を担う主体として、「AI 開発者」、「AI 提供者」及び「AI 利用者」の 3 つに大別され、それぞれ以下のとおり定義される。これらの主体は事業者（又は各者内の部

[2] 一般消費者が AI、特に生成 AI を活用するにあたっての注意点等については、消費者庁「AI 利活用ハンドブック～生成 AI 編～」（2024 年 5 月）を参照。
https://www.caa.go.jp/policies/policy/consumer_policy/information/ai_handbook
また、消費者を支援することに活用できる AI を含むデジタル技術の現状や見通し、課題等については、内閣府「消費者をエンパワーするデジタル技術に関する専門調査会」報告書で整理されている。
https://www.cao.go.jp/consumer/iinkaikouhyou/2024/doc/202412_digital_technology_houkoku.pdf

門）を想定しており、AI の活用方法によっては同一の事業者が AI 開発者、AI 提供者又は AI 利用者の複数を兼ねる場合もある（「図 3. 一般的な AI 活用の流れにおける主体の対応」参照）[3]。

- **AI 開発者（AI Developer）**
  AI システムを開発する事業者（AI を研究開発する事業者を含む）
  AI モデル・アルゴリズムの開発、データ収集（購入を含む）、前処理、AI モデル学習及び検証を通して AI モデル、AI モデルのシステム基盤、入出力機能等を含む AI システムを構築する役割を担う。

- **AI 提供者（AI Provider）**
  AI システムをアプリケーション、製品、既存のシステム、ビジネスプロセス等に組み込んだサービスとして AI 利用者（AI Business User）、場合によっては業務外利用者に提供する事業者
  AI システム検証、AI システムの他システムとの連携の実装、AI システム・サービスの提供、正常稼働のための AI システムにおける AI 利用者（AI Business User）側の運用サポート又は AI サービスの運用自体を担う。AI サービスの提供に伴い、様々なステークホルダーとのコミュニケーションが求められることもある。

- **AI 利用者（AI Business User）**
  事業活動において、AI システム又は AI サービスを利用する事業者
  AI 提供者が意図している適正な利用を行い、環境変化等の情報を AI 提供者と共有し正常稼働を継続すること又は必要に応じて提供された AI システムを運用する役割を担う。また、AI の活用において業務外利用者に何らかの影響が考えられる場合[4]は、当該者に対する AI による意図しない不利益の回避、AI による便益最大化の実現に努める役割を担う。

[3] 開発・提供 ・利用の対象に生成 AI も含まれる。AI 提供者又は AI 利用者が政府・自治体等、公的機関になる場合は、民間事業者の場合とは別の考えが必要になる可能性がある。

[4] 業務外利用者は、AI 利用者の指示及び注意に従わない場合、何らかの被害を受ける可能性があることを留意する必要がある。

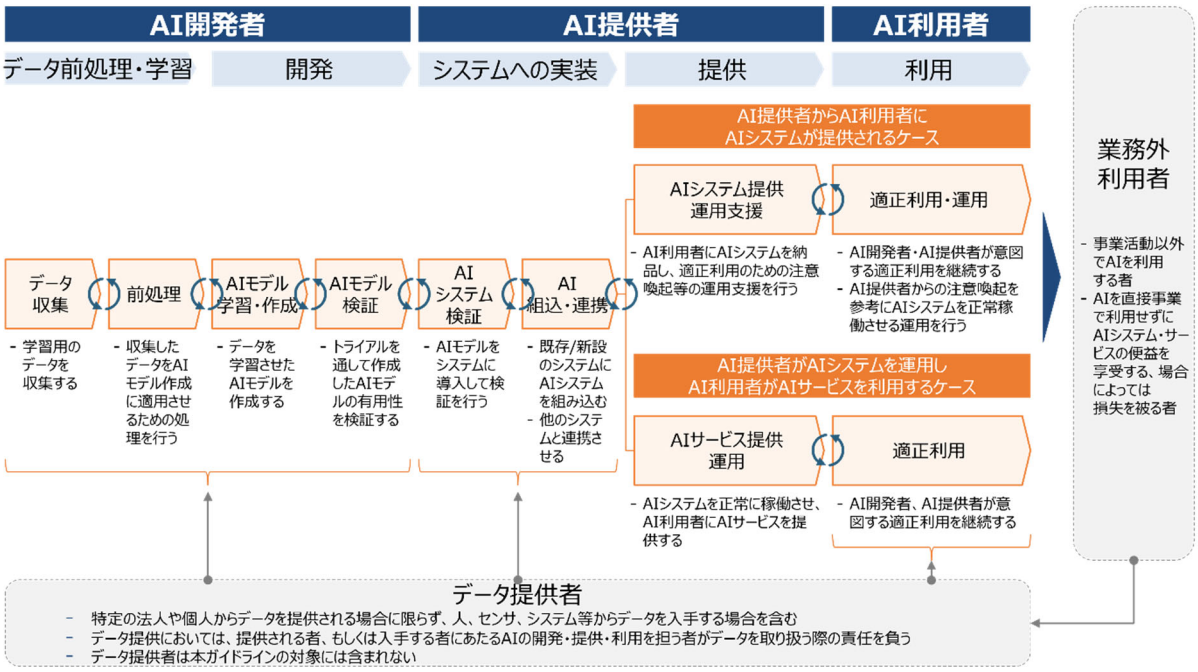

図 3. 一般的な AI 活用の流れにおける主体の対応

自らが該当する「AI 開発者」、「AI 提供者」又は「AI 利用者」の立場から、「ステークホルダーからの期待を鑑みつつどのような社会を目指すのか（「基本理念」= why）」を踏まえ、「AI に関しどのような取組を行うべきか（指針 = what）」を明らかにすることが重要であり、また指針を実現するために、「具体的にどのようなアプローチで取り組むか（実践 = how）」を検討・決定し、実践することが AI の安全安心な活用に有用と考えられる。実際の AI システム・サービスは目的・活用技術・データ・利用環境等によって多様なユースケースとなり、技術の発展等、外部環境の変化も踏まえつつ、AI 開発者、AI 提供者及び AI 利用者が連携して最適なアプローチを検討することが重要である。本ガイドラインは読みやすさを考慮し、本編で「基本理念」及び「指針」を扱い、別添（付属資料）で「実践」を扱うこととする。

「基本理念」及び「指針」を扱う本ガイドラインの本編の構成を以下に記載する。

- **第 1 部**

  本ガイドラインの内容に関する理解を助けるために、「用語の定義」を中心に記載する。

- **第 2 部**

  AI の活用により目指すべき社会及びそれを実現するための「基本理念」（why）、並びに原則及び各主体に共通する指針（what）を記載する。AI の活用による便益を求める中で、AI が社会にリスクをもたらす可能性を鑑み、「共通の指針」を実践するために必要となるガバナンスの構築についても触れる。第 2 部では第 3 部以降のもととなる内容を解説しているため、AI を活用する全ての事業者が内容を確認し、理解することが重要である。

- **第 3 部～第 5 部**

  AI を活用した事業活動を担う 3 つの主体に関し、第 2 部では触れられない主体毎の留意点を記載する。AI を活用する事業者は自らに関する事項を理解することが重要であり、それと同時に、隣接する主

体と関係する事項が多く存在するため、当該主体以外に関する事項も理解することが重要である（「図 4. 本ガイドラインの構成」参照）。

「AI 開発者」、「AI 提供者」及び「AI 利用者」においては、第 1 部、第 2 部に加えて、第 3 部以降の当該部及び別添（付属資料）を確認することで、AI を活用する際のリスク及びその対応方針の基本的な考え方を把握することが可能となる。特に具体的な取組が決まっていない事業者にとっては、別添の記載例が参考となるため、別添の関連箇所を中心とした確認が重要となる。なお、経営者を含む事業執行責任者[5]は、その職務を全うするために、本ガイドラインにおける「基本理念」（why）及び指針（what）を踏まえて、事業戦略と一体で AI を活用する際のリスク対策を検討・実践し、AI の安全安心な活用を推進することが重要である。

図 4. 本ガイドラインの構成

AI をめぐる環境はグローバル規模で日進月歩の進化を続けていることから、AI を活用する事業者は、国際的な動向にも注意を払うことが重要である。我が国においても、こうした現状を踏まえ、広島 AI プロセス[6]を通じて、AI に関する国際的な共通理解及び指針の策定を主導しており、2023 年 12 月には広島 AI プロセス包括的政策枠組み等を取りまとめた[7]。本ガイドラインも同プロセスへの貢献を意図するとともに、同プロセスを含む国際

[5] 事業執行責任者は、政府・自治体等の公的機関の事業執行責任者も含む。

[6] 2023 年 5 月の G7 広島サミットの結果を受けて、生成 AI に関する国際的なルールの検討を行うため、「広島 AI プロセス」を立ち上げた。その後、同年 9 月の「広島 AI プロセス閣僚級会合」、10 月の京都 IGF での「マルチステークホルダーハイレベル会合」等を経て発出された「広島 AI プロセスに関する G7 首脳声明」を踏まえ、同年 12 月に「G7 デジタル・技術大臣会合」を開催し、同年の成果として、「広島 AI プロセス包括的政策枠組み」を取りまとめた。さらに、2024 年には、3 月の「G7 産業・技術・デジタル大臣会合」、10 月の「G7 デジタル・技術大臣会合」を経て、国際行動規範の遵守状況に係る「報告枠組み」が、同年 12 月に G7 で合意された。https://www.soumu.go.jp/hiroshimaaiprocess/

[7] GPAI（The Global Partnership on Artificial Intelligence）とは、人間中心の考え方に立ち、「責任ある AI」の開発・利用をプロジェクトベースの取組で推進するために設立（2020 年）された、政府・国際機関・産業界・有識者・市民社会等のマルチステークホルダーによる国際連携イニシアティブである。2024 年 7 月、GPAI は OECD との統合パートナーシップ体制へと移行した。GPAI 専門家による調査研究やプロジェクトに対

的な議論を踏まえながら検討したものである。一方で、AI をめぐる考え方及び法令は国・地域で異なることから、特に、国境を越えた活動を行う事業者は、現地の法令に対応すべきであり、ステークホルダーの要望に応じた対応が期待される。特に高度な AI システムについては市場に導入される前の安全性評価[8]の枠組みを検討する等、国・地域によってはガバナンスの実効性を担保するための措置を講じている場合もあり、注意を払うことが重要である。[9]

し、広島 AI プロセスの推進にも協力している生成 AI に関するプロジェクトをはじめとした運営・管理面での支援を提供するため、2024 年 7 月に「GPAI 東京専門家支援センター」を国立研究開発法人情報通信研究機構(NICT)に設置。同センターでは、これまでに SAFE プロジェクト（生成 AI の安全性を保証するための実践的なアプローチの調査）を展開してきたほか、OECD と協議しながら新たな生成 AI 関連プロジェクトを実施していくこととしている。
https://www2.nict.go.jp/gpai-tokyo-esc/

[8] 2023 年 11 月、英国では、高度な AI システムの評価の開発、実施等を行う「AI Safety Institute」の設置計画を発表し、米国では、米国標準技術研究所（NIST）に AI リスクマネジメントフレームワークの実装、レッドチーミングの評価等を行う「US AI Safety Institute」を設立することを発表した。なお、英国では、2025 年 2 月に、AI Safety Institute から AI Security Institute に改称されている。日本においても、これらの海外機関と連携しつつ、AI の開発・提供・利用の安全性向上に資する基準・ガイダンス等の検討、AI の安全性評価方法等の調査、AI の安全性に関する技術・事例の調査等を目的とし、2024 年 2 月 14 日に、関係府省庁及び関係機関の協力の下、「AI セーフティ・インスティテュート」を独立行政法人情報処理推進機構(IPA)に設置。本ガイドラインと米国 NIST AI リスクマネジメントフレームワーク（RMF）のクロスウォークや「AI セーフティに関する評価観点ガイド」（2024 年 9 月）
https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety/の公表、各国の AISI 等との連携・意見交換、各種調査等の活動を実施している。
https://aisi.go.jp/

[9] 2025 年 2 月 4 日に、AI 戦略会議・AI 制度研究会 中間とりまとめが公表された。
https://www8.cao.go.jp/cstp/ai/interim_report.pdf

第１部 AIとは

AIはArtificial Intelligence（人工知能）を意味し、1956年にダートマス会議で初めて使用された言葉であるとされている。AIは未だ確立された定義は存在しないが、「人工」・「知能」とあるように、人間の思考プロセスと同じような形で動作するコンピュータープログラム、コンピューター上で知的判断を下せるシステム等を指す。機械学習（Machine Learning、略してML）を行わない、専門家の知識を大量にインプットすることで知識にもとづく推論を行うエキスパートシステムと呼ばれるものも、AIとみなす考えもある。しかし、2000年代以降、ディープラーニング等による「画像認識」、「自然言語処理（翻訳等）」、「音声認識」が活用されるようになり、特定の分野に特化し、予測、提案又は決定を行うことができるシステムをAIと指すようになってきた。また、2021年以降、基盤モデル[10]の台頭により、特定の分野のみに特化したAIではない、汎用的なAIの開発が進んでいる。その結果、「予測」、「提案」、「決定」にとどまらず、画像、文章等を生成する「生成AI」が普及するようになり、注目を集めている。このように、ひとくくりに「AI」と言っても、その種類は多岐にわたり、今後のAI技術の在り方については専門家であっても予測することは困難である。

このような状況を踏まえつつ、本ガイドラインにおける関連する用語を以下のとおり定義する。

関連する用語

- **AI**
  現時点で確立された定義はなく（統合イノベーション戦略推進会議決定「人間中心のAI社会原則」（2019年3月29日））、広義の人工知能の外延を厳密に定義することは困難である。本ガイドラインにおけるAIは「AIシステム（以下に定義）」自体又は機械学習をするソフトウェア若しくはプログラムを含む抽象的な概念とする。
  （参考としてJIS X 22989:2023ではISO/IEC 22989:2022にもとづき、以下のように定義されている）
  ＜学問分野＞AIシステムのメカニズム及び適用の研究開発
  注釈1. 研究開発は、コンピュータサイエンス、データサイエンス、自然科学、人文科学、数学等、幾つもの分野にわたって行うことが可能である

- **AIシステム**
  活用の過程を通じて様々なレベルの自律性をもって動作し学習する機能を有するソフトウェアを要素として含むシステムとする（機械、ロボット、クラウドシステム等）。
  （参考としてJIS X 22989:2023ではISO/IEC 22989:2022にもとづき、以下のように定義されている）
  人間が定義した所与の目標の集合に対して、コンテンツ、予測、推奨、意思決定等の出力を生成する工学的システム
  注釈1. 工学的システムは、人工知能に関連する様々な技法及びアプローチを使用して、作業の実施

[10] 大規模言語モデルに代表される基盤モデルは、様々なサービスを支える個別モデルを生み出すコアの技術基盤である。基盤モデルから派生する下流の幅広いタスクに適応させたモデルの開発、開発過程そのものから得られる知見等の観点から、一般的なAIとは異なる性質を持つ。

に使用可能であるデータ、知識、プロセス等を表すモデルを開発することが可能である
注釈 2. AI システムは、様々な自動化のレベルで動作するように設計されている
（参考として OECD AI Principles overview では以下のように定義されている）
AI システムは、明示的又は暗黙的な目的のために推測するマシンベースのシステムである。受け取った入力から、物理環境又は仮想環境に影響を与える可能性のある予測、コンテンツ、推奨、意思決定等の出力を生成する。AI システムが異なれば、導入後の自律性及び適応性のレベルも異なる

- **高度な AI システム**
  最先端の基盤モデル及び生成 AI システムを含む、最も高度な AI システムを指す。
  （広島 AI プロセスでの定義を引用）

- **AI モデル（ML モデル）**
  AI システムに含まれ、学習データを用いた機械学習によって得られるモデルで、入力データに応じた予測結果を生成する。
  （参考として JIS X 22989:2023 では ISO/IEC 22989:2022 にもとづき、以下のように定義されている）
  入力データ又は情報にもとづいて推論（inference）又は予測を生成する数学的構造
  例：単変量線形関数 $y=\theta_0+\theta_1 x$ が、線形回帰を使用して訓練されている場合、結果のモデルは、$y=3+7x$ のようになる
  注釈 1. 機械学習モデルは、機械学習アルゴリズムにもとづく訓練の結果として得られる

- **AI サービス**
  AI システムを用いた役務を指す。AI 利用者への価値提供の全般を指しており、AI サービスの提供・運営は、AI システムの構成技術に限らず、人間によるモニタリング、ステークホルダーとの適切なコミュニケーション等の非技術的アプローチも連携した形で実施される。

- **生成 AI**
  文章、画像、プログラム等を生成できる AI モデルにもとづく AI の総称を指す。

- **AI ガバナンス**
  AI の利活用によって生じる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、そこからもたらされる正のインパクト（便益）を最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計並びに運用。

第２部 AI により目指すべき社会及び各主体が取り組む事項

第 2 部では、まず、AI により目指す社会としての「A. 基本理念」を記載する。更に、その実現に向け、各主体が取り組む「B. 原則」とともに、そこから導き出される「C. 共通の指針」を記載する。また、高度な AI システムに関係する事業者が遵守すべき「D. 高度な AI システムに関係する事業者に共通の指針」を記載する。加えて、この「C. 共通の指針」を実践し AI を安全安心に活用していくために重要な「E. AI ガバナンスの構築」についても記載する。

A. 基本理念

「はじめに」で述べたとおり、我が国が 2019 年 3 月に策定した「人間中心の AI 社会原則」においては、AI が Society 5.0 の実現に貢献することが期待されている。また、AI を人類の公共財として活用し、社会の在り方の質的変化及び真のイノベーションを通じて地球規模の持続可能性とつなげることが重要であることが述べられている。そして、以下の 3 つの価値を「基本理念」として尊重し、「その実現を追求する社会を構築していくべき」としている。

① 人間の尊厳が尊重される社会（Dignity）

AI を利活用して効率性や利便性を追求するあまり、人間が AI に過度に依存したり、人間の行動をコントロールすることに AI が利用される社会を構築するのではなく、人間が AI を道具として使いこなすことによって、人間の様々な能力をさらに発揮することを可能とし、より大きな創造性を発揮したり、やりがいのある仕事に従事したりすることで、物質的にも精神的にも豊かな生活を送ることができるような、人間の尊厳が尊重される社会を構築する必要がある

② 多様な背景を持つ人々が多様な幸せを追求できる社会（Diversity and Inclusion）

多様な背景、価値観又は考え方を持つ人々が多様な幸せを追求し、それらを柔軟に包摂した上で新たな価値を創造できる社会は、現代における一つの理想であり、大きなチャレンジである。AI という強力な技術は、この理想に我々を近づける一つの有力な道具となりうる。我々は AI の適正な開発と展開によって、このように社会の在り方を変革していく必要がある

③ 持続可能な社会（Sustainability）

我々は、AI の活用によりビジネスやソリューションを次々と生み、社会の格差を解消し、地球規模の環境問題や気候変動等にも対応が可能な持続性のある社会を構築する方向へ展開させる必要がある。科学・技術立国としての我が国は、その科学的・技術的蓄積を AI によって強化し、そのような社会を作ることに貢献する責務がある

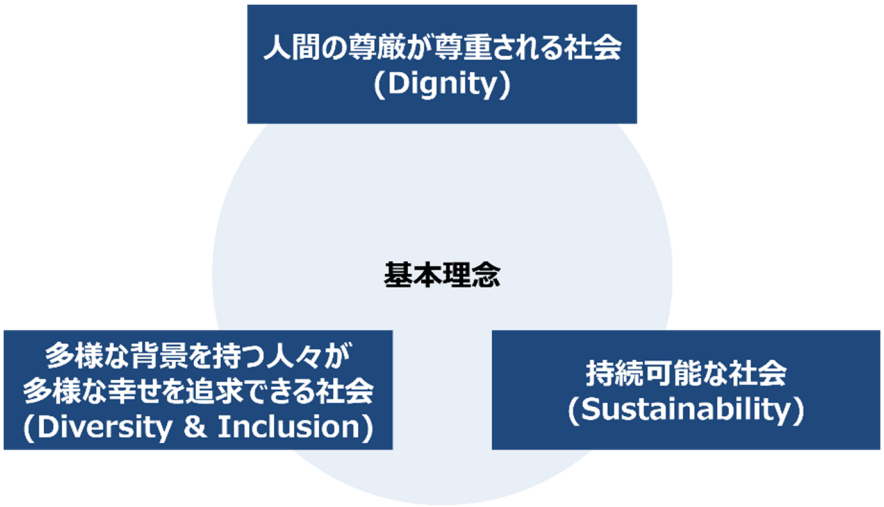

図 5. 基本理念

この基本的な考え方自体は、著しい技術の発展によっても変わるものではなく、目指すべき理念であり続けている。したがって、AI の発展に伴い、日本及び多国間の枠組みで目指すべき方向性として、これらの「基本理念」が尊重されるべきである。

B. 原則

「基本理念」を実現するためには、各主体がこれに沿う形で取組を進めることが重要であり、そのために各主体が念頭におく「原則」を、各主体が取り組む事項及び社会と連携した取組が期待される事項に整理した。この「原則」は、「人間中心の AI 社会原則」を土台としつつ、OECD の AI 原則等の海外の諸原則を踏まえ、再構成したものである。

各主体が取り組む事項

各主体は、「基本理念」より導き出される人間中心の考え方をもとに[11]、AI システム・サービスの開発・提供・利用を促進し、人間の尊厳を守りながら、事業における価値の創出、社会課題の解決等、AI の目的を実現していくことが重要である。このため、各主体は、AI 活用に伴う社会的リスクの低減を図るべく、安全性・公平性といった価値を確保することが重要である。また、個人情報の不適正な利用等の防止を始めとするプライバシー保護、並びに AI システムの脆弱性等による可用性の低下及び外部からの攻撃等のリスクに対応するセキュリティ確保を行うことが重要である。これらを実現するために、各主体は、システムの検証可能性を確保しながらステークホルダー[12]に対する適切な情報を提供することにより透明性を向上させ、アカウンタビリティを果たすことが重要となる。

[11] 下線部は、後段の「C.共通の指針」として整理されるものである。

[12] ステークホルダー：AI 開発者、AI 提供者、AI 利用者及び業務外利用者以外の第三者を含む AI の活用によって直接・間接の影響を受ける可能性がある全ての主体（以降同様）

加えて、今後、AI アーキテクチャの多様化に伴うバリューチェーン変動等により、各主体の役割が変動する可能性を踏まえた上で、各主体間で連携し、バリューチェーン全体での AI の品質の向上に努めること及びマルチステークホルダーで継続して議論していくことが重要である。

このような対応を行うことで、各主体が、AI のリスクを最低限に抑制しつつ、AI システム・サービスの開発・提供・利用を通じて最大限の便益を享受することが期待される。

社会と連携した取組が期待される事項

AI による社会への便益を一層増大させ、我々が目指すべき「基本理念」を実現していくためには、各主体それぞれの取組に加え、社会（政府・自治体及びコミュニティも含む）と積極的に連携することが期待される。このため、各主体は、社会と連携して、社会の分断を回避し、全ての人々に AI の恩恵が行き渡るための教育・リテラシー確保の機会を提供することが期待される。加えて、新たなビジネス・サービスが創出され、持続的な経済成長の維持及び社会課題の解決策が提示されるよう、公正競争の確保及びイノベーションの促進に貢献していくことが期待される。

C. 共通の指針

取組にあたり、各主体は、以下に述べる「1） 人間中心」に照らし、法の支配、人権、民主主義、多様性・包摂性及び公平公正な社会を尊重するよう AI システム・サービスを開発・提供・利用すべきである。また、憲法、知的財産関連法令及び個人情報保護法をはじめとする関連法令、AI に係る個別分野の既存法令等を遵守すべきであり、国際的な指針等の検討状況についても留意することが重要である[13]。

なお、これらの取組は、各主体が開発・提供・利用する AI システム・サービスの特性、用途、目的及び社会的文脈を踏まえ、各主体の資源制約を考慮しながら自主的に進めることが重要である。

各主体が連携して、バリューチェーン全体で取り組むべきことは、具体的には、以下のとおり整理される。

1) 人間中心

各主体は、AI システム・サービスの開発・提供・利用において、後述する各事項を含む全ての取り組むべき事項が導出される土台として、少なくとも憲法が保障する又は国際的に認められた人権を侵すことがないようにすべきである。また、AI が人々の能力を拡張し、多様な人々の多様な幸せ（well-being）の追求が可能となるように行動することが重要である。

① 人間の尊厳及び個人の自律

- ✧ AI が活用される際の社会的文脈を踏まえ、人間の尊厳及び個人の自律を尊重する
- ✧ 特に、AI を人間の脳・身体と連携させる場合には、その周辺技術に関する情報を踏まえつつ、諸外国及び研究機関における生命倫理の議論等を参照する

[13] 事業の地理的な展開状況、開発された AI モデルを用いる AI 提供者・AI 利用者の所在地、学習を行うサーバの所在地等に応じ、各準拠法に従う必要がある。我が国の国内法に準拠する場合は、データの類型に応じ、個人情報、知的財産権等にそれぞれ適用される法令に適合した取扱を行う。また、データの取扱においては、法令で定められていなくともステークホルダー間の契約関係において、利用が禁止される場合が存在することにも留意すべきである。

✧ 個人の権利・利益に重要な影響を及ぼす可能性のある分野において AI を利用したプロファイリングを行う場合、個人の尊厳を尊重し、アウトプットの正確性を可能な限り維持させつつ、AI の予測、推奨、判断等の限界を理解して利用し、かつ生じうる不利益等を慎重に検討した上で、不適切な目的に利用しない

② AI による意思決定・感情の操作等への留意

✧ 人間の意思決定、認知等、感情を不当に操作することを目的とした、又は意識的に知覚できないレベルでの操作を前提とした AI システム・サービスの開発・提供・利用は行わない

✧ AI システム・サービスの開発・提供・利用において、自動化バイアス[14]等の AI に過度に依存するリスクに注意を払い、必要な対策を講じる[15]

✧ フィルターバブル[16]に代表されるような情報又は価値観の傾斜を助長し、AI 利用者を含む人間が本来得られるべき選択肢が不本意に制限されるような AI の活用にも注意を払う

✧ 特に、選挙、コミュニティでの意思決定等をはじめとする社会に重大な影響を与える手続きに関連しうる場合においては、AI の出力について慎重に取り扱う

③ 偽情報等への対策

✧ 生成 AI によって、内容が真実・公平であるかのように装った情報を誰でも作ることができるようになり、AI が生成した偽情報・誤情報・偏向情報が社会を不安定化・混乱させるリスクが高まっていることを認識した上で、必要な対策を講じる[17,18]

④ 多様性・包摂性の確保

✧ 公平性の確保に加え、いわゆる「情報弱者」及び「技術弱者」を生じさせず、より多くの人々が AI の恩恵を享受できるよう社会的弱者による AI の活用を容易にするよう注意を払う

- ユニバーサルデザイン、アクセシビリティの確保、関連するステークホルダー[19]への教育・フォローアップ 等

⑤ 利用者支援

---

[14] 人間の判断や意思決定において、自動化されたシステムや技術への過度の信頼や依存が生じる現象を指す。

[15] 必要な対策としては、AI による評価、判断、推奨、又は予測等のアウトプットを人間が鵜呑みにしないために、AI の欠点を含む特性を理解できるトレーニングを受けることや、AI の評価や判断等を人間が承認する際には人間自身が AI の評価や判断等を承認する理由や根拠を独自に考えてから承認すべきこと等が提案されている。

[16] 「フィルターバブル」とは、アルゴリズムがネット利用者個人の検索履歴やクリック履歴を分析し学習することで、個々にとっては望むと望まざるとにかかわらず見たい情報が優先的に表示され、利用者の観点に合わない情報からは隔離され、自身の考え方や価値観の「バブル（泡）」の中に孤立するという情報環境を指す。このようなもともとある人間の傾向とネットメディアの特性の相互作用による現象と言われているものとして、「フィルターバブル」の他、「エコーチェンバー」も挙げられる。そういったリスクがある一方で、AI はパーソナライズされた的を絞った返答を AI 利用者や業務外利用者に提供し、有益な形で提案を行うことを可能とするという便益もある。

[17] インターネット上の偽・誤情報の流通・拡散への対応を含む、デジタル空間における情報流通の健全性確保に向けた今後の対応方針と具体的な方策について検討するため、「デジタル空間における情報流通の健全性確保の在り方に関する検討会」が総務省において開催され、2024 年 9 月に「とりまとめ」が公表された。（https://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000417.html）
また、技術的対応として、総務省により「インターネット上の偽・誤情報等対策技術の開発・実証事業」が実施されている。
（https://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000415.html）

[18] RAG（検索拡張生成）の活用等により、ハルシネーションの抑制や出力過程・根拠の透明性向上等が期待されている。

[19] 関連するステークホルダー：AI 開発者、AI 提供者、AI 利用者及び業務外利用者を含む直接・間接問わず AI の活用に関与する主体（以降同様）

✧ 合理的な範囲で、AI システム・サービスの機能及びその周辺技術に関する情報を提供し、選択の機会の判断のための情報を適時かつ適切に提供する機能が利用可能である状態とする
- デフォルトの設定、理解しやすい選択肢の提示、フィードバックの提供、緊急時の警告、エラーへの対処等

⑥ 持続可能性の確保

✧ AI システム・サービスの開発・提供・利用において、ライフサイクル全体で、地球環境への影響も検討する（特に、生成 AI など計算量の多い AI システムにおいては、モデルの軽量化や目的に応じたモデルの使い分け等の対策を講じる）

これら全てを前提とした上で、各主体は、AI のパフォーマンス（有用性）を可能な範囲で高め、人々に便益及び豊かさを与え、幸福を実現することが期待される。

2) 安全性

各主体は、AI システム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすべきである。加えて、精神及び環境に危害を及ぼすことがないようにすることが重要である。

① 人間の生命・身体・財産、精神及び環境への配慮

✧ AI システム・サービスの出力の正確性を含め、要求に対して十分に動作している（信頼性）

✧ 様々な状況下でパフォーマンスレベルを維持し、無関係な事象に対して著しく誤った判断を発生させないようにする（堅牢性（robustness））

✧ AI の活用又は意図しない AI の動作によって生じうる権利侵害の重大性、侵害発生の可能性等、当該 AI の性質・用途等に照らし、必要に応じて定期的かつ客観的なモニタリング及び対処も含めて人間がコントロールできる制御可能性を確保する

✧ 適切なリスク分析を実施し、リスクへの対策（回避、低減、移転又は容認）を講じる

✧ 人間の生命・身体・財産、精神及び環境へ危害を及ぼす可能性がある場合は、講ずべき措置について事前に整理し、ステークホルダーに関連する情報を提供する
- 関連するステークホルダーが講ずべき措置及び利用規則を明記する

✧ AI システム・サービスの安全性を損なう事態が生じた場合の対処方法を検討し、当該事態が生じた場合に速やかに実施できるよう整える

② 適正利用

✧ 主体のコントロールが及ぶ範囲で本来の目的を逸脱した提供・利用[20]により危害が発生することを避けるべく、AI システム・サービスの開発・提供・利用を行う

✧ マルチモーダルな生成 AI を中心とする AI システム・サービスの生成物については、比較的容易により精巧な生成物の生成が可能となるため、その精巧さにより誤解や偏見等を助長する可能

[20] なお、RAG（検索拡張生成）を活用した場合には生成 AI による回答の収束が加速する可能性が高いため、例えばコンテンツの多様性・独創性を必要とする業務については、RAG（検索拡張生成）の活用が適切ではない場合もあることに留意が必要である。

性があることや、他者の知的財産権等を侵害する可能性があること等にも留意しつつ、その利用に際し人間の判断を介在させる等[21]の対策を講じる[22]

③ 適正学習[23]

- ✧ AI システム・サービスの特性及び用途を踏まえ、学習等に用いるデータの正確性・必要な場合には最新性（データが適切であること）等を確保する
- ✧ 学習等に用いるデータの透明性の確保、法的枠組みの遵守、AI モデルの更新等を合理的な範囲で適切に実施する
- ✧ 権利侵害複製物等の違法なコンテンツ[24]や情報等を学習データに含めないこと等に留意する[25]

3) 公平性

各主体は、AI システム・サービスの開発・提供・利用において、特定の個人ないし集団への人種、性別、国籍、年齢、政治的信念、宗教等の多様な背景を理由とした不当で有害な偏見及び差別をなくすよう努めることが重要である。また、各主体は、それでも回避できないバイアスがあることを認識しつつ、この回避できないバイアスが人権及び多様な文化を尊重する観点から許容可能か評価した上で、AI システム・サービスの開発・提供・利用を行うことが重要である。[26]

① AI モデルの各構成技術に含まれるバイアスへの配慮

---

[21] 人間の判断を介在させるための工夫としては、生成物であることを示す透かしを入れることやユーザーインタフェースの改善等があげられる。

[22] なお、プログラムコードを生成する AI を活用する場合は、生成したコードの安全性やセキュリティ面への配慮が重要となる。

[23] AI 提供者・AI 利用者においてもファインチューニング、再学習を行う場合は AI 開発者と同様に安全性の担保に努めることが重要となる。

[24] 文化庁「AI と著作権に関する考え方について」（文化審議会著作権分科会法制度小委員会、2024 年 3 月）において、海賊版等、違法にアップロードされているものも学習されてしまうことが権利者の懸念として挙げられている。

[25] 知的財産関連法令との関係については内閣府、文化庁等での議論が進められており、今後の検討状況についても留意すべきである。特に AI と著作権に関する考え方については、文化審議会著作権分科会法制度小委員会にて取りまとめており、各主体においては、その趣旨を踏まえた対応方針を検討することが重要となる。

・文化庁「AI と著作権に関する考え方について」（文化審議会著作権分科会法制度小委員会、2024 年 3 月）
https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf
・内閣府「AI 時代の知的財産権検討会 中間とりまとめ」（知的財産戦略推進事務局、2024 年 5 月）
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf
・文化庁「AI と著作権に関するチェックリスト＆ガイダンス」（文化庁著作権課、2024 年 7 月）
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_01.pdf
・内閣府「AI 時代の知的財産権検討会 中間とりまとめ 手引き（権利者向け）」（知的財産戦略推進事務局、2024 年 11 月）
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/2411_tebiki.pdf
なお、「AI 時代の知的財産権検討会 中間とりまとめ」及び「AI と著作権に関するチェックリスト＆ガイダンス」では、本ガイドラインと異なり、「AI 開発者」、「AI 提供者」及び「AI 利用者」に加えて、「業務外利用者（一般利用者）」及び「権利者」も対象にして、各主体に期待される取組事項例等を整理している。

[26] 「バイアス」という言葉には例えば以下の通り様々な解釈が考えられ、本ガイドラインではそれらを総称したものとして用いている。
・統計的な用語（サンプリングバイアス。偏り・偏差など。）
・心理学的な用語（認知バイアス（思い込み等に起因。集団ごとの社会通念等による社会的バイアスも含む）、感情バイアス（人間の感情・都合等に起因）など。）
また NIST の Proposal for Identifying and Managing Bias in Artificial Intelligence (SP 1270) では、AI において、Systemic（既存のルールや規範、慣行等によるもの）、Statistical and Computational (統計的・計数的なもの)、Human（認知・知覚、習性等によるもの）という 3 つのカテゴリと、それぞれに属する典型的なバイアスがある旨が説明されている。

- ✧ バイアスを生み出す要因は多岐に渡るため、各技術要素（学習データ、AI モデルの学習過程、AI 利用者又は業務外利用者が入力するプロンプト[27]、AI モデルの推論時に参照する情報、連携する外部サービス等）及び AI 利用者の振る舞いを含めて、公平性の問題となりうるバイアスの要因となるポイントを特定する
- ✧ AI システム・サービスの特性又は用途によっては、ステークホルダーが意図しない又は感知できないバイアスが生じる可能性についても検討する

② 人間の判断の介在

- ✧ AI の出力結果が公平性を欠くことがないよう、AI に単独で判断させるだけでなく、適切なタイミングで人間の判断を介在させる利用を検討する。なおその際には、人間の判断が自動化バイアスに左右されないような対策[28]を講じるべきである
- ✧ バイアスが生じていないか、AI システム・サービスの目的、制約、要件及び決定を明確かつ透明性のある方法により分析し、対処するためのプロセスを導入する
- ✧ ステークホルダーが意図しない又は感知できないバイアスや、潜在的なバイアス（学習データには表れないバイアス）[29]に留意し、多様な背景、文化又は分野のステークホルダーと対話した上で、方針を決定する

4) プライバシー保護

各主体は、AI システム・サービスの開発・提供・利用において、その重要性に応じ、プライバシーを尊重し、保護することが重要である。その際、関係法令を遵守すべきである。

① AI システム・サービス全般におけるプライバシーの保護

- ✧ 個人情報保護法[30]等の関連法令の遵守、各主体のプライバシーポリシーの策定・公表等により、社会的文脈及び人々の合理的な期待を踏まえ、ステークホルダーのプライバシーが尊重され、保護されるよう、その重要性に応じた対応を取る
- ✧ 以下の事項を考慮しつつ、プライバシー保護のための対応策を検討する
  - 個人情報保護法にもとづいた対応の確保
  - 国際的な個人データ保護の原則及び基準の参照[31]

---

[27] 大規模言語モデルを始めとする生成 AI では、AI 利用者は、コンテキスト内学習と呼ばれる学習方法により、学習済パラメータを更新することなく、AI 利用者の入力（プロンプトと呼ばれる）に応じて、特定のタスクに対する学習を行わせることが可能である。

[28] 必要な対策については、前掲脚注 15 を参照。

[29] 潜在的なバイアスの例として、人種や性別等の明らかなセンシティブ属性を取り除いたデータであっても、関連する情報からセンシティブ属性が類推されてしまう場合や、複数の属性の組み合わせによってバイアスが生じてしまう場合（交差バイアス）などがある。

[30] 個人情報保護法については、個人情報保護委員会において、いわゆる 3 年ごと見直しに関する検討が進められている。その中で、AI 開発等を含む統計作成等のみを目的とした取り扱いを実施する場合の本人同意の在り方等が検討されている。
https://www.ppc.go.jp/personalinfo/3nengotominaoshi/
（2025 年 2 月公表資料：https://www.ppc.go.jp/files/pdf/seidotekikadainitaisurukangaekatanitsuite_r6.pdf）

[31] OECD, Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, OECD/LEGAL/0188, ISO/IEC 29100:2011 Information technology Security techniques Privacy framework 等プライバシーに関する国際的な指針を踏まえることが期待される。また、より広範囲での個人データの円滑な越境移転及び各国における規律の相互運用性を促進させる等の目的で Global Cross-Border Privacy Rules （CBPR） Forum が立ち上がっており、日本も 2022 年 4 月に参加し、Global CBPR Framework を公表している。また、生成 AI に関しては、G7 データ保護プライバシー機関ラウンドテーブル会合による「生成 AI に関する声明」（2023 年 6 月）及び GPA（Global Privacy Assembly）による「生成 AI システムに関する決議」（2023 年 10 月）も参照。

5) セキュリティ確保

各主体は、AI システム・サービスの開発・提供・利用において、不正操作によって AI の振る舞いに意図せぬ変更又は停止が生じることのないように、セキュリティを確保することが重要である。

① AI システム・サービスに影響するセキュリティ対策[32]

- ✧ AI システム・サービスの機密性・完全性・可用性を維持し、常時、AI の安全安心な活用を確保するため、その時点での技術水準に照らして合理的な対策を講じる
- ✧ AI システム・サービスの特性を理解し、正常な稼働に必要なシステム間の接続が適切に行われているかを検討する
- ✧ 推論対象データに微細な情報を混入させることで関連するステークホルダーの意図しない判断が行われる可能性を踏まえて、AI システム・サービスの脆弱性を完全に排除することはできないことを認識する

② 最新動向への留意

- ✧ AI システム・サービスに対する外部からの攻撃は日々新たな手法が生まれており、これらのリスクに対応するための留意事項を確認する

6) 透明性[33]

各主体は、AI システム・サービスの開発・提供・利用において、AI システム・サービスを活用する際の社会的文脈を踏まえ、AI システム・サービスの検証可能性を確保しながら、必要かつ技術的に可能な範囲で、ステークホルダーに対し合理的な範囲で情報を提供することが重要である。

① 検証可能性の確保

- ✧ AI の判断にかかわる検証可能性を確保するため、データ量又はデータ内容に照らし合理的な範囲で、AI システム・サービスの開発過程、利用時の入出力等、AI の学習プロセス、推論過程、判断根拠等のログを記録・保存する
- ✧ ログの記録・保存にあたっては、利用する技術の特性及び用途に照らして、事故等の原因究明、再発防止策の検討、損害賠償責任要件の立証上の重要性等を踏まえて、記録方法、

[32] 詳細な手法については、英国サイバーセキュリティセンター（NCSC）「セキュアな AI システム開発のためのガイドライン （Guidelines for secure AI system development） 」（2023 年 11 月）が公開されている。
https://www8.cao.go.jp/cstp/stmain/20231128ai.html

[33] 透明性については、諸外国でも様々な定義がある。例えば、NIST,Artificial Intelligence Risk Management Framework（January 2023）では、透明性（システムで何が起きたかについて答えられること）、説明可能性（システムでどのように決定がなされたかについて答えられること）及び解釈可能性（なぜその決定がされたかについてその意味又は文脈について答えられること）に分類されており、European Commission, ETHICS GUIDELINES FOR TRUSTWORTHY AI （April 2019）では、トレーサビリティ、説明可能性及びコミュニケーションが取り上げられている。また、国際標準(ISO/IEC JTC1/SC42)では、透明性（適切な情報が関係者に提供されること）と定義されている。その他に EU, AI 法（Artificial Intelligence Act）（June 2024）において透明性とは、AI システムが適切な追跡可能性と説明可能性を実現する方法で開発及び利用され、人間が AI システムで通信又は対話していることを認識できるようにし、AI システムの機能と限界について利用者に適切に開示し、影響を受ける人間に対して権利について説明することを意味する。本文書では、情報開示に関する事項を広く「透明性」とする。なお定義のほか、開示対象者や開示主体、開示目的が諸外国によって異なることにも留意する。

頻度、保存期間等について検討する

② 関連するステークホルダーへの情報提供

- ✧ AI との関係の仕方、AI の性質、目的等に照らして、それぞれが有する知識及び能力に応じ、例えば、以下について取りまとめた情報の提供及び説明を行う
  - AI システム・サービス全般
    - AI を利用しているという事実及び活用している範囲
    - データ収集及びアノテーションの手法
    - 学習及び評価の手法
    - 基盤としている AI モデルに関する情報
    - AI システム・サービスの能力、限界及び提供先における適正/不適正な利用方法
    - AI システム・サービスの提供先、AI 利用者が所在する国・地域等において適用される関連法令等
- ✧ 多様なステークホルダーとの対話を通じて積極的な関与を促し、社会的な影響及び安全性に関する様々な意見を収集する
- ✧ 加えて、実態に即して、AI システム・サービスを提供・利用することの優位性、それに伴うリスク等を関連するステークホルダーに示す

③ 合理的かつ誠実な対応

- ✧ 上記の「②関連するステークホルダーへの情報提供」は、アルゴリズム又はソースコードの開示を必ずしも想定するものではなく、プライバシー及び営業秘密を尊重して、採用する技術の特性及び用途に照らし、社会的合理性が認められる範囲で実施する（ただし、後掲する「④関連するステークホルダーへの説明可能性・解釈可能性の向上」や『７） アカウンタビリティ』の要請を満たす必要がある）
- ✧ 公開されている技術を用いる際には、それぞれ定められている規程に準拠する
- ✧ 開発した AI システムのオープンソース化にあたっても、社会的な影響を検討する

④ 関連するステークホルダーへの説明可能性・解釈可能性の向上

- ✧ 関連するステークホルダーの納得感及び安心感の獲得、また、そのための AI の動作に対する証拠の提示等を目的として、説明する主体がどのような説明が求められるかを分析・把握できるよう、説明を受ける主体がどのような説明が必要かを共有し、必要な対応を講じる
  - AI 提供者：AI 開発者に、どのような説明が必要となるかを共有する
  - AI 利用者：AI 開発者・AI 提供者に、どのような説明が必要となるかを共有する

7) アカウンタビリティ[34]

各主体は、AI システム・サービスの開発・提供・利用において、トレーサビリティの確保、「共通の指針」の対応状況等について、ステークホルダーに対して、各主体の役割及び開発・提供・利用する AI システム・サ

[34] アカウンタビリティを説明可能性と定義することもあるが、本ガイドラインでは情報開示は透明性で対応することとし、アカウンタビリティとは AI に関する事実上・法律上の責任を負うこと及びその責任を負うための前提条件の整備に関する概念とする。

ービスのもたらすリスクの程度を踏まえ、合理的な範囲でアカウンタビリティを果たすことが重要である。

① トレーサビリティの向上

- ✧ データの出所、AI システム・サービスの開発・提供・利用中に行われた意思決定等について、技術的に可能かつ合理的な範囲で追跡・遡求が可能な状態を確保する

② 「共通の指針」の対応状況の説明

- ✧ 「共通の指針」の対応状況について、ステークホルダー（サプライヤーを含む）に対してそれぞれが有する知識及び能力に応じ、例えば以下の事項を取りまとめた情報の提供及び説明を定期的に行う
  - 全般
    - 「共通の指針」の実践を妨げるリスクの有無及び程度に関する評価
    - 「共通の指針」の実践の進捗状況
  - 「人間中心」関連
    - 偽情報等への留意、多様性・包摂性、利用者支援及び持続可能性の確保の対応状況
  - 「安全性」関連
    - AI システム・サービスに関する既知のリスク及び対応策、並びに安全性確保の仕組み
  - 「公平性」関連
    - AI モデルを構成する各技術要素（学習データ、AI モデルの学習過程、AI 利用者又は業務外利用者が入力すると想定するプロンプト、AI モデルの推論時に参照する情報、連携する外部サービス等）によってバイアスが含まれうること
  - 「プライバシー保護」関連
    - AI システム・サービスにより自己又はステークホルダーのプライバシーが侵害されるリスク及び対応策、並びにプライバシー侵害が発生した場合に講ずることが期待される措置
  - 「セキュリティ確保」関連
    - AI システム・サービスの相互間連携又は他システムとの連携が発生する場合、その促進のために必要な標準準拠等
    - AI システム・サービスがインターネットを通じて他の AI システム・サービス等と連携する場合に発生しうるリスク及びその対応策

③ 責任者の明示

- ✧ 各主体においてアカウンタビリティを果たす責任者を設定する

④ 関係者間の責任の分配

- ✧ 関係者間の責任について、業務外利用者も含めた主体間の契約、社会的な約束（ボランタリーコミットメント）等により、責任の所在を明確化する

⑤ ステークホルダーへの具体的な対応

- ✧ 必要に応じ、AI システム・サービスの利用に伴うリスク管理、安全性確保のための各主体の AI ガバナンスに関するポリシー、プライバシーポリシー等の方針を策定し、公表する（社会及び一般市民に対するビジョンの共有、並びに情報発信・提供を行うといった社会的責任を含む）
- ✧ 必要に応じ、AI の出力の誤り等について、ステークホルダーからの指摘を受け付ける機会を設けるとともに、定期的かつ客観的なモニタリングを実施する
- ✧ ステークホルダーの利益を損なう事態が生じた場合[35]、どのように対応するか方針を策定してこれを着実に実施し、進捗状況については必要に応じて定期的にステークホルダーに報告する

⑥ 文書化[36]

- ✧ 上記に関する情報を文書化して一定期間保管し、必要なときに、必要なところで、入手可能かつ利用に適した形で参照可能な状態とする

各主体が社会と連携して取り組むことが期待される事項は、具体的には、下記のとおり整理される。

8) 教育・リテラシー

各主体は、主体内の AI に関わる者が、AI の正しい理解及び社会的に正しい利用ができる知識・リテラシー・倫理感を持つために、必要な教育を行うことが期待される。また、各主体は、AI の複雑性、誤情報といった特性及び意図的な悪用の可能性もあることを勘案して、ステークホルダーに対しても教育を行うことが期待される。[37]

① AI リテラシーの確保

- ✧ 各主体内の AI に関わる者が、その関わりにおいて十分なレベルの AI リテラシーを確保するために必要な措置を講じる

② 教育・リスキリング

---

[35] 特に、学習データ等の権利者や AI の評価、判断、奨励、又は予測等によって不利益を被った者等から問合せや情報開示の要望を受けた場合には、必要な範囲で適切な対応を行うことが望ましい。また、裁判所からの開示命令等があった場合には、当該命令等に従って必要な手続きを取ることが求められる。

[36] 「文書化」については、後から容易に確認可能となるよう適切なツールで記録が残されていれば差し支えなく、必ずしも紙媒体や特定の文書の形式による必要はない。

[37] 経済産業省・IPA は、個人の学習及び企業の人材確保・育成の指針として DX 時代の人材像を「デジタルスキル標準」として整理（2022 年 12 月）。2023 年 8 月に「生成 AI 時代の DX 推進に必要な人材・スキルの考え方」を取りまとめ、指示（プロンプト）の習熟、「問いを立てる」「仮説検証する」等の必要性をスキル標準に反映し、2024 年 7 月には、普及する生成 AI の影響を踏まえ、新技術への向き合い方等を補記として追加し、「データ活用」「テクノロジー」の学習項目例に「大規模言語モデル・画像生成モデル・オーディオ生成モデル」等の生成 AI 関連の技術を追加。また、経済産業省は、知的財産権等の権利・利益の保護に十分に配慮した、コンテンツ制作における生成 AI の適切な利活用の方向性を示す「コンテンツ制作のための生成 AI 利活用ガイドブック」を公表（2024 年 7 月）。
https://www.meti.go.jp/policy/mono_info_service/contents/ai_guidebook_set.pdf
・デジタルスキル標準　https://www.meti.go.jp/policy/it_policy/jinzai/skill_standard/main.html
・デジタル時代の人材政策に関する検討会　https://www.meti.go.jp/shingikai/mono_info_service/digital_jinzai/index.html
また、総務省は今後の生活の中で生成 AI に触れうる国民（初心者）向けに、生成 AI の基礎知識、生成 AI の活用場面や入門的な使い方、生成 AI 活用時の注意点などを紹介する「生成 AI はじめの一歩～生成 AI の入門的な使い方と注意点～」を公表。
https://www.soumu.go.jp/use_the_internet_wisely/special/generativeai/

- ✧ 生成 AI の活用拡大によって、AI と人間の作業の棲み分けが変わっていくと想定されるため、新たな働き方ができるよう教育・リスキリング等を検討する[38]
- ✧ 様々な人が AI で得られる便益の理解を深め、リスクに対するレジリエンスを高められるよう、世代間ギャップも考慮した上での教育の機会を提供する

③ ステークホルダーへのフォローアップ

- ✧ AI システム・サービス全体の安全性を高めるため、必要に応じて、ステークホルダーに対して教育及びリテラシー向上のためのフォローアップを行う

9) 公正競争確保

各主体は、AI を活用した新たなビジネス・サービスが創出され、持続的な経済成長の維持及び社会課題の解決策の提示がなされるよう、AI をめぐる公正な競争環境の維持に努めることが期待される。

10) イノベーション

各主体は、社会全体のイノベーションの促進に貢献するよう努めることが期待される。

① オープンイノベーション等の推進

- ✧ 国際化・多様化、産学官連携及びオープンイノベーションを推進する
- ✧ AI のイノベーションに必要なデータが創出される環境の維持に配慮する

② 相互接続性・相互運用性への留意

- ✧ 自らの AI システム・サービスと他の AI システム・サービスとの相互接続性及び相互運用性を確保する
- ✧ 標準仕様がある場合には、それに準拠する

③ 適切な情報提供

- ✧ 自らのイノベーションを損なわない範囲で必要な情報提供を行う

以上の事項に加え、AI 開発者、AI 提供者又は AI 利用者のそれぞれで重要となる事項は、「表 1. 共通の指針に加えて主体毎に重要となる事項」のとおり整理される。なお、表の「-」が記載されている箇所は、各主体による第 2 部 C.「共通の指針」記載の事項にもとづく対応が期待されており、対応不要を意味するものではない。

なお、特に AI と知的財産権との関係については、AI 開発者、AI 提供者又は AI 利用者の各主体に期待される取組につき、内閣府「AI 時代の知的財産権検討会 中間とりまとめ」（2024 年 5 月）や文化庁著作権課「AI と著作権に関するチェックリスト＆ガイダンス」（2024 年 7 月）において整理されている。各主体においては、その趣旨を踏まえた対応方針を検討することが重要となる。

[38] 厚生労働省の雇用政策研究会にて「新たなテクノロジー等を活用した労働生産性の向上」というテーマで、技術変化を踏まえたキャリア形成やスキル教育について触れられている。
https://www.mhlw.go.jp/stf/shingi/other-syokuan_128950.html

以降は「表 1.「共通の指針」に加えて主体毎に重要となる事項」に記載されている内容（項目）を、[主体 – 指針番号） 記載内容.]のルールにて識別・表記する。

- 主体は、AI 開発者（AI **D**eveloper）、AI 提供者（AI **P**rovider）及び AI 利用者（AI Business **U**ser）の頭文字を用い、指針番号及び記載内容の番号は同表に記載の番号にて表記する

例）D-2） i. は AI 開発者の安全性に関する適切なデータの学習についての重要事項を指す

表 1.「共通の指針」に加えて主体毎に重要となる事項

	第 2 部. C.共通の指針	「共通の指針」に加えて主体毎に重要となる事項		
		第 3 部. AI 開発者 (D)	第 4 部. AI 提供者 (P)	第 5 部. AI 利用者 (U)
1) 人間中心	① 人間の尊厳及び個人の自律 ② AI による意思決定・感情の操作等への留意 ③ 偽情報等への対策 ④ 多様性・包摂性の確保 ⑤ 利用者支援 ⑥ 持続可能性の確保	-	-	-
2) 安全性	① 人間の生命・身体・財産、精神及び環境への配慮 ② 適正利用 ③ 適正学習	ⅰ. 適切なデータの学習 ⅱ. 人間の生命・身体・財産、精神及び環境に配慮した開発 ⅲ. 適正利用に資する開発	ⅰ. 人間の生命・身体・財産、精神及び環境に配慮したリスク対策 ⅱ. 適正利用に資する提供	ⅰ. 安全を考慮した適正利用
3) 公平性	① AI モデルの各構成技術に含まれるバイアスへの配慮 ② 人間の判断の介在	ⅰ. データに含まれるバイアスへの配慮 ⅱ. AI モデルのアルゴリズム等に含まれるバイアスへの配慮	ⅰ. AI システム・サービスの構成及びデータに含まれるバイアスへの配慮	ⅰ. 入力データ又はプロンプトに含まれるバイアスへの配慮
4) プライバシー保護	① AI システム・サービス全般におけるプライバシーの保護	ⅰ. 適切なデータの学習 (D-2) ⅰ. 再掲)	ⅰ. プライバシー保護のための仕組み及び対策の導入 ⅱ. プライバシー侵害への対策	ⅰ. 個人情報の不適切入力及びプライバシー侵害への対策
5) セキュリティ確保	① AI システム・サービスに影響するセキュリティ対策 ② 最新動向への留意	ⅰ. セキュリティ対策のための仕組みの導入 ⅱ. 最新動向への留意	ⅰ. セキュリティ対策のための仕組みの導入 ⅱ. 脆弱性への対応	ⅰ. セキュリティ対策の実施
6) 透明性	① 検証可能性の確保 ② 関連するステークホルダーへの情報提供 ③ 合理的かつ誠実な対応 ④ 関連するステークホルダーへの説明可能性・解釈可能性の向上	ⅰ. 検証可能性の確保 ⅱ. 関連するステークホルダーへの情報提供	ⅰ. システムアーキテクチャ等の文書化 ⅱ. 関連するステークホルダーへの情報提供	ⅰ. 関連するステークホルダーへの情報提供
7) アカウンタビリティ	① トレーサビリティの向上 ②「共通の指針」の対応状況の説明 ③ 責任者の明示 ④ 関係者間の責任の分配 ⑤ ステークホルダーへの具体的な対応 ⑥ 文書化	ⅰ. AI 提供者への「共通の指針」の対応状況の説明 ⅱ. 開発関連情報の文書化	ⅰ. AI 利用者への「共通の指針」の対応状況の説明 ⅱ. サービス規約等の文書化	ⅰ. 関連するステークホルダーへの説明 ⅱ. 提供された文書の活用及び規約の遵守
8) 教育・リテラシー	① AI リテラシーの確保 ② 教育・リスキリング ③ ステークホルダーへのフォローアップ	-	-	-
9) 公正競争確保	-	-	-	-
10) イノベーション	① オープンイノベーション等の推進 ② 相互接続性・相互運用性への留意 ③ 適切な情報提供	ⅰ. イノベーションの機会創造への貢献	-	-

D. 高度な AI システムに関係する事業者に共通の指針

高度な AI システムに関係する事業者は、広島 AI プロセスを経て策定された「全ての AI 関係者向けの広島プロセス国際指針」及びその基礎となる「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指針」を踏まえ、「共通の指針」に加え、以下を遵守すべきである[39]。ただし、I） ～ XI）は高度な AI システムを開発する AI 開発者にのみ適用される内容もあるため、第 3～5 部に後述のとおり、AI 提供者及び AI 利用者は適切な範囲で遵守することが求められる。

I) AI ライフサイクル全体にわたるリスクを特定、評価、軽減するために、高度な AI システムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる（「2） 安全性」、「6） 透明性」「7） アカウンタビリティ」）

- 具体的には、レッドチーム[40]等の様々な手法を組み合わせて、多様/独立した内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緩和策を実施する
- 上記テストを支援するために、開発中に行われた意思決定に関するトレーサビリティを確保するように努める

II) 市場投入を含む導入後、脆弱性、及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する（「5） セキュリティ確保」、「7） アカウンタビリティ 」）

- リスクレベルに見合った適切なタイミングで、AI システムの活用状況のモニタリングを実施し、それらに対処するための適切な措置を講じる
  - ✧ 他の利害関係者と協力して、報告されたインシデントの適切な文書化を維持し、特定されたリスクと脆弱性を軽減することが奨励される

III) 高度な AI システムの能力、限界、適切・不適切な使用領域を公表し、十分な透明性の確保を支援することで、アカウンタビリティの向上に貢献する（「6） 透明性」、「7） アカウンタビリティ 」）

- データの出所に始まり、どのような意思決定を行ったかについて、合理的な説明を行い、トレーサビリティを確保するため文書化・公表する
- 関連するステークホルダーが AI システムの出力を解釈し、AI 利用者や業務外利用者が適切に活用できるようにするために、明確で理解可能な形で文書化・公表する

IV) 産業界、政府、市民社会、学界を含む、高度な AI システムを開発する組織間での責任ある情報共有とインシデントの報告に向けて取り組む（「5） セキュリティ確保」、「6） 透明性」、「7）アカウンタビリティ」、「10）イノベーション」）

---

[39] 詳細は、G7 デジタル・技術大臣会合（2023 年 12 月）で採択された「広島 AI プロセス G7 デジタル・技術閣僚声明」における「広島 AI プロセス包括的政策枠組み」の「II. 全ての AI 関係者向け及び高度な AI システムを開発する組織向けの広島プロセス国際指針」を参照。
https://www.soumu.go.jp/menu_news/s-news/01tsushin06_02000283.html

[40] 攻撃者がどのように対象組織を攻撃するかの観点で、セキュリティへの対応体制及び対策の有効性を確認するチームを意味する。なお、AISI からレッドチーミング手法ガイドが公表されている。
AISI「AI セーフティに関するレッドチーミング手法ガイド」（2024 年 9 月）
https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/

- 具体的には、モニタリング結果の報告書やセキュリティや安全性のリスクに関する関連文書等が含まれる

V) 特に高度な AI システム開発者に向けた、個人情報保護方針及び緩和策を含む、リスクベースのアプローチにもとづく AI ガバナンス及びリスク管理方針を策定し、実施し、開示する（「4） プライバシー保護」、「7）アカウンタビリティ」）
- 適切な場合には、プライバシーポリシーを公表する
- AI ガバナンスに関するポリシーや実行するための組織を確立し、開示することが期待される

VI) AI のライフサイクル全体にわたり、物理的セキュリティ、サイバーセキュリティ、内部脅威に対する安全対策を含む、強固なセキュリティ管理に投資し、実施する（「5） セキュリティ確保」）
- 情報セキュリティのための運用上の対策や適切なサイバー/物理的アクセス制御等も検討する

VII) 技術的に可能な場合は、電子透かしやその他の技術等、AI 利用者及び業務外利用者が、AI が生成したコンテンツを識別できるようにするための、信頼できるコンテンツ認証及び来歴のメカニズムを開発し、導入する（「6）透明性」）
- 具体的には、適切かつ技術的に実現可能な場合、組織の高度な AI システムで作成されたコンテンツ認証及び来歴メカニズムが含まれる
- 透かし等を通じた特定のコンテンツが高度な AI システムで作成されたかどうかを AI 利用者及び業務外利用者が判断できるツールや API の開発に努める
  - ✧ AI 利用者及び業務外利用者が AI システムと相互作用していることを知ることができるよう、ラベリングや免責事項の表示等、その他の仕組みを導入することが奨励される

VIII) 社会的、安全、セキュリティ上のリスクを軽減するための研究を優先し、効果的な軽減策への投資を優先する（「10） イノベーション」）
- AI の安全性、セキュリティ、信頼性の向上やリスクへの対処に関する研究が含まれる

IX) 世界の最大の課題、特に気候危機、世界保健、教育等（ただしこれらに限定されない）に対処するため、高度な AI システムの開発を優先する（「10） イノベーション」）
- 信頼性のある人間中心の AI 開発に向けた取組を実施し、同時に業務外利用者も含めたリテラシーの向上のための支援をする

X) 国際的な技術規格の開発を推進し、適切な場合にはその採用を推進する（「10） イノベーション」）
- 電子透かしを含む国際的な技術標準とベストプラクティスの開発に貢献し、適切な場合にはそれを利用し、標準開発組織（SDO）と協力する

XI) 適切なデータインプット対策を実施し、個人データ及び知的財産を保護する（「2） 安全性」、「3） 公平性」）
- 有害な偏見バイアスを軽減するために、訓練データやデータ収集等、データの質を管理するための適切な措置を講じることが奨励される
- 訓練用データセットの適切な透明性も支援されるべきであり、適用される法的枠組みを遵守する

XII) 高度な AI システムの信頼でき責任ある利用を促進し、貢献する（「5） セキュリティ確保」、「8） 教育・リテラシー」）

- 高度な AI システムが特定のリスク（例えば偽情報の拡散に関するもの）をどのように増大させるか、新たなリスクをどのように生み出すか等の課題を含め、各主体及びステークホルダーのリテラシーや認識等の向上のための機会を提供する
- 各主体間で連携し、高度な AI システムに関する新たなリスクや脆弱性を特定し、対処するための情報共有を行うことが奨励される

E. AI ガバナンスの構築

各主体間で連携しバリューチェーン全体で「共通の指針」を実践し AI を安全安心に活用していくためには、AI に関する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、そこからもたらされる便益を最大化するための、AI ガバナンスの構築が重要となる。また、「Society 5.0」を実現するためには、サイバー空間とフィジカル空間を高度に融合させたシステム（CPS）の社会実装を進めつつ、その適切な AI ガバナンスを構築することが不可欠である。CPS を基盤とする社会は、複雑で変化が速く、リスクの統制が困難であり、こうした社会の変化に応じて、AI ガバナンスが目指すゴールも常に変化していく。そのため、事前にルール又は手続が固定された AI ガバナンスではなく、企業・法規制・インフラ・市場・社会規範といった様々なガバナンスシステムにおいて、「環境・リスク分析」「ゴール設定」「システムデザイン」「運用」「評価」といったサイクルを、マルチステークホルダーで継続的かつ高速に回転させていく、「アジャイル・ガバナンス」の実践が重要となる[41]。

なお、具体的な検討にあたっては開発・提供・利用予定の AI のもたらすリスクの程度及び蓋然性、並びに各主体の資源制約に配慮することが重要である。

① まず、AI システム・サービスがライフサイクル全体においてもたらしうる便益/リスク、開発・運用に関する社会的受容、「外部環境の変化」、AI 習熟度等を踏まえ、対象となる AI システム・サービスに関連する「環境・リスク分析」を実施する

② これを踏まえ、AI システム・サービスを開発・提供・利用するか否かを判断し、開発・提供・利用する場合には、AI ガバナンスに関するポリシーの策定等を通じて「AI ガバナンス・ゴール[42]の設定」を検討する。なお、この AI ガバナンス・ゴールは、各主体の存在意義、理念・ビジョンといった経営上のゴールと整合したものとなるように設定する

[41] 別添（付属資料）において、経済産業省「AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1」を土台とした、AI ガバナンス実践のための詳細解説に加え、各主体の具体的な取組事項としての「行動目標」及び各主体を想定した仮想的な「実践例」も記載しているため、ご参照のこと。

[42] AI ガバナンス・ゴールとして、本ガイドラインに記載の「共通の指針」への対応事項からなる自社の取組方針（「AI ポリシー」等、呼称は各主体により相違）及び「共通の指針」への対応事項を包含しつつそれ以外の要素を含む取組方針（データ活用ポリシー等）を設定すること等が考えられる。AI を活用することによって多様性・包摂性を向上させる等の便益を高めるための指針を提示してもよい。また、呼称も各主体に委ねられている。

③ 更に、この AI ガバナンス・ゴールを達成するための「AI マネジメントシステムの設計」を行った上で、これを「運用」する。その際には、各主体が、AI ガバナンス・ゴール及びその運用状況について外部の「ステークホルダーに対する透明性、アカウンタビリティ（公平性等）」を果たすようにする

④ その上で、リスクアセスメント等をはじめとして、AI マネジメントシステムが有効に機能しているかを継続的にモニタリングし、「評価」及び継続的改善を実施する

⑤ AI システム・サービスの運用開始後も、規制等の社会的制度の変更等の「外部環境の変化」を踏まえ、再び「環境・リスク分析」を実施し、必要に応じてゴールを見直す

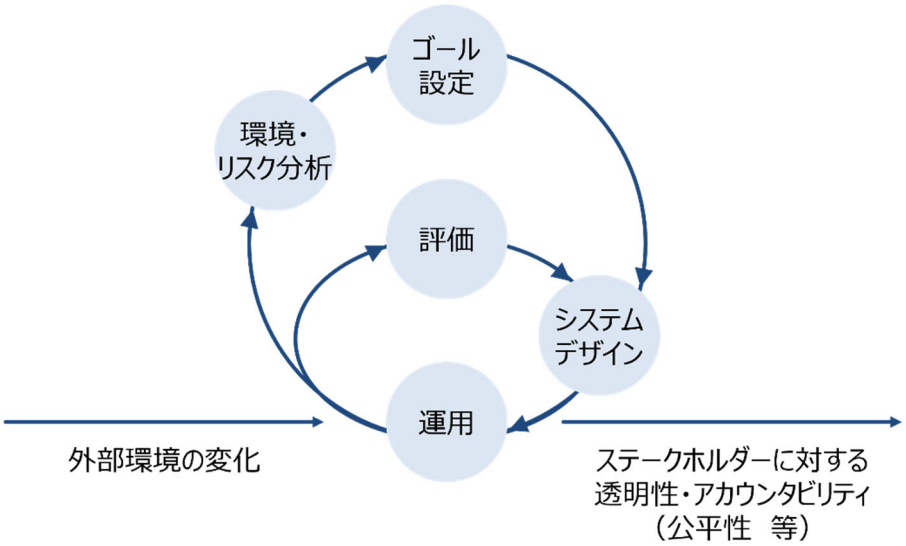

図 6. アジャイル・ガバナンスの基本的なモデル

また、AI ガバナンスの検討にあたっては、バリューチェーンを念頭に置き、以下の点に留意することが重要である。

- バリューチェーン/リスクチェーンの観点で主体間の連携を確保する
  - 複数主体にまたがる論点の例：AI リスク把握、品質の向上、各 AI システム・サービスが相互に繋がること（System of Systems）による新たな価値の創出、AI 利用者又は業務外利用者のリテラシー向上等
  - 主体間で整理が必要になりうる点の例：学習及び利用に用いるデータ・生成された AI モデルに関する権利関係の契約等
- データの流通をはじめとしたリスクチェーンの明確化、並びに開発・提供・利用の各段階に適したリスク管理及び AI ガバナンス体制の構築を実施する
  - AI 開発からサービス実施にわたるバリューチェーン/リスクチェーンが複数国にまたがることが想定される場合、データの自由な流通（Data Free Flow with Trust、以下、「DFFT」という）の確保のための適切な AI ガバナンスに係る国際社会の検討状況の把握及びそれを踏まえた相互運用性の確保（「標準」及び「枠組み間の相互運用性」の二側面）

これらを効果的な取組とするためには、経営層の責任は大きく、そのため、リーダーシップを発揮することが重要である。その際は、短期的な利益の追求の観点から AI ガバナンスの構築を単なるコストと捉えるのではなく、各

主体の持続的成長及び中長期的な発展を志向した先行投資として捉えることが重要である。そのリーダーシップの下、上記のアジャイル・ガバナンスのサイクルを回しつつ、各組織の戦略及び企業体制に AI ガバナンスを落とし込んでいくことで、各組織の中で文化として根付かせることが期待される。

第３部 AI 開発者に関する事項

AI 開発者は、AI モデルを直接的に設計し変更を加えることができるため、AI システム・サービス全体においても AI の出力に与える影響力が高い。また、イノベーションを牽引することが社会から期待され、社会全体に与える影響が非常に大きい。このため、自身の開発する AI が提供・利用された際にどのような影響を与えるか、事前に可能な限り検討し、対応策を講じておくことが重要となる。

AI 開発の現場においては、時に、正確性を重視するためにプライバシー又は公平性が損なわれたり、プライバシーを過度に重視して透明性が損なわれたり等、リスク同士又は倫理観の衝突の場面がある。その場合、当該事業者における経営リスク及び社会的な影響力を踏まえ、適宜判断・修正していくことが重要である。また、AI システムにおいて予期せぬ事故が発生した際に、AI のバリューチェーンに連なる者は、何らかの説明を求められる立場に立つ可能性があることを念頭に置き、AI 開発者としても、どのような関与を行ったかについて、合理的な説明を行うことができるよう記録を残すことが重要である[43]。

以下に AI 開発者にとって重要な事項を挙げる。

- **データ前処理・学習時**

  - D-2）i. 適切なデータの学習
    - ✧ プライバシー・バイ・デザイン等を通じて、学習時のデータについて、適正に収集するとともに、第三者の個人情報、知的財産権に留意が必要なもの等が含まれている場合には、法令に従って適切に扱うことを、AI のライフサイクル全体を通じて確保する（「2）安全性」、「4）プライバシー保護」、「5）セキュリティ確保」）
    - ✧ 学習前・学習全体を通じて、データのアクセスを管理するデータ管理・制限機能の導入検討を行う等、適切な保護措置を実施する（「2）安全性」、「5）セキュリティ確保」）

  - D-3）i. データに含まれるバイアスへの配慮
    - ✧ 学習データ、AI モデルの学習過程によってバイアス（学習データには現れない潜在的なバイアスを含む）が含まれうることに留意し、データの質を管理するための相当の措置を講じる（「3）公平性」）
    - ✧ 学習データ、AI モデルの学習過程からバイアスを完全に排除できないことを踏まえ、AI モデルが代表的なデータセットで学習され、AI システムに不公正なバイアス（特定の個人・集団に対し、合理的に説明できない不利益をもたらすバイアス）がないか点検されることを確保する（「3）公平性」）

- **AI 開発時**

  - D-2）ii. 人間の生命・身体・財産、精神及び環境に配慮した開発

[43] AI の開発等にあたり、AI の信頼性を向上させるためのツールや指標のカタログが、OECD により提供されている。
https://oecd.ai/en/catalogue/overview

- ✧ ステークホルダーの生命・身体・財産、精神及び環境に危害を及ぼすことがないよう、以下の事項を検討する（「2）安全性」）
  - 様々な状況下で予想される利用条件下でのパフォーマンスだけでなく、予期しない環境での利用にも耐えうる性能の要求
  - リスク（連動するロボットの制御不能、不適切な出力等）を最小限に抑える方法の要求（ガードレール技術等）

➢ D-2）iii. 適正利用に資する開発

- ✧ 開発時に想定していない AI の提供・利用により危害が発生することを避けるため、安全に利用可能な AI の使い方について明確な方針・ガイダンスを設定する（「2）安全性」）
- ✧ 事前学習済の AI モデルに対する事後学習を行う場合に、学習済 AI モデルを適切に選択する（商用利用可能なライセンスかどうか、事前学習データ、学習・実行に必要なスペック等）（「2）安全性」）

➢ D-3）ii. AI モデルのアルゴリズム等に含まれるバイアスへの配慮

- ✧ AI モデルを構成する各技術要素（AI 利用者又は業務外利用者が入力するプロンプト、AI モデルの推論時に参照する情報、連携する外部サービス等）によってバイアスが含まれうることまで検討する（「3）公平性」）
- ✧ AI モデルからバイアスを完全に排除できないことを踏まえ、AI モデルが代表的なデータセットで学習され、AI システムに不公正なバイアスがないか点検されることを確保する（「3）公平性」）

➢ D-5）i. セキュリティ対策のための仕組みの導入

- ✧ AI システムの開発の過程を通じて、採用する技術の特性に照らし適切にセキュリティ対策を講ずる（セキュリティ・バイ・デザイン）（「5）セキュリティ確保」）

➢ D-6）i. 検証可能性の確保

- ✧ AI の予測性能及び出力の品質が、活用開始後に大きく変動する可能性又は想定する精度に達しないこともある特性を踏まえ、事後検証のための作業記録を保存しつつ、その品質の維持・向上を行う（「2）安全性」、「6）透明性」）

- **AI 開発後**

➢ D-5）ii. 最新動向への留意

- ✧ AI システムに対する攻撃手法は日々新たなものが生まれており、これらのリスクに対応するため、開発の各工程で留意すべき点を確認する[44]（「5）セキュリティ確保」）

➢ D-6）ii. 関連するステークホルダーへの情報提供

[44] IPA「AI（Artificial Intelligence）の推進」（https://www.ipa.go.jp/digital/ai/index.html）等を通じて情報を収集することができる。

- ✧ 自らの開発する AI システムについて、例えば以下の事項を適時かつ適切に関連するステークホルダーに（AI 提供者を通じて行う場合を含む）情報を提供する（「6）透明性」）
  - AI システムの学習等による出力又はプログラムの変化の可能性（「1）人間中心」）
  - AI システムの技術的特性、安全性確保の仕組み、利用の結果生じる可能性のある予見可能なリスク及びその緩和策等の安全性に関する情報（「2）安全性」）
  - 開発時に想定していない AI の提供・利用により危害が発生することを避けるための AI 開発者が意図する利用範囲（「2）安全性」）
  - AI システムの動作状況に関する情報並びに不具合の原因及び対応状況（「2）安全性」）
  - AI の更新を行った場合の内容及びその理由の情報（「2）安全性」）
  - AI モデルで学習するデータの収集ポリシー、学習方法及び実施体制等（「3）公平性」、「4）プライバシー保護」、「5）セキュリティ確保」）

- ➢ D-7）i. AI 提供者への「共通の指針」の対応状況の説明
  - ✧ AI 提供者に対して、AI には活用開始後に予測性能又は出力の品質が大きく変動する可能性、想定する精度に達しないこともある旨、その結果生じうるリスク等の情報提供及び説明を行う。具体的には以下の事項を周知する（「7）アカウンタビリティ」）
    - AI モデルを構成する各技術要素（学習データ、AI モデルの学習過程、AI 利用者又は業務外利用者が入力すると想定するプロンプト、AI モデルの推論時に参照する情報、連携する外部サービス等）において含まれる可能性があるバイアスへの対応等（「3）公平性」）

- ➢ D-7）ii. 開発関連情報の文書化
  - ✧ トレーサビリティ及び透明性の向上のため、AI システムの開発過程、意思決定に影響を与えるデータ収集及びラベリング、使用されたアルゴリズム等について、可能な限り第三者が検証できるような形で文書化する（「7）アカウンタビリティ」）
    （注） ここで文書化されたものをすべて開示するという意味ではない

以下に、AI 開発者の取組が期待される事項を挙げる。

- ➢ D-10）i. イノベーションの機会創造への貢献
  - ✧ 可能な範囲で以下の事項を実施し、イノベーションの機会の創造に貢献することが期待される（「10）イノベーション」）
    - AI の品質・信頼性、開発の方法論等の研究開発を行う
    - 持続的な経済成長の維持及び社会課題の解決策が提示されるよう貢献する
    - DFFT 等の国際議論の動向の参照、AI 開発者コミュニティ又は学会への参加の取組等、国際化・多様化及び産学官連携を推進する
    - 社会全体への AI に関する情報提供を行う

「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」における追加的な記載事項

高度な AI システムを開発する AI 開発者については、上記に加え、「第 2 部 D. 高度な AI システムに関係する事業者に共通の指針」及び「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」[45]を遵守すべきである。

以下、「第 2 部 D. 高度な AI システムに関係する事業者に共通の指針」との比較において、当該「行動規範」において追加的に記載されている事項を示す。なお、当該「行動規範」全体の内容については、「別添 3. C. 高度な AI システムの開発にあたって遵守すべき事項」を参照のこと。

I. AI ライフサイクル全体にわたるリスクを特定、評価、軽減するために、高度な AI システムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる
   - リスク軽減のための緩和策を文書化するとともに、定期的に更新すべき。また、各主体はセクターを超えた関係者と連携してこれらのリスクへの緩和策を評価し、採用すべき

II. 市場投入を含む導入後、脆弱性、及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する
   - 報奨金制度、コンテスト、賞品等を通じて、責任を持って弱点を開示するインセンティブを与えることの検討を奨励

III. 高度な AI システムの能力、限界、適切・不適切な使用領域を公表し、十分な透明性の確保を支援することで、アカウンタビリティの向上に貢献する
   - 透明性報告書は、使用説明書や関連する技術的文書等とともに、最新に保たれるべき

IV. 産業界、政府、市民社会、学界を含む、高度な AI システムを開発する組織間での責任ある情報共有とインシデントの報告に向けて取り組む
   - AI システムの安全性やセキュリティ等を確保するための共有の基準、メカニズムを開発、推進すべき。加えて、AI のライフサイクル全体にわたって、適切な文書化や他の主体との協力、関連情報の共有や社会への報告を実施すべき

V. 特に高度な AI システム開発者に向けた、個人情報保護方針及び緩和策を含む、リスクベースのアプローチにもとづく AI ガバナンス及びリスク管理方針を策定し、実施し、開示する
   - 可能であれば、AI のライフサイクル全体を通じて、AI リスクを特定・評価・予防・対処するための AI ガバナンス方針を策定・実施・開示し、定期的に更新すべき。また、事業者の職員等に対する教育方針を確立すべき

VI. AI のライフサイクル全体にわたり、物理的セキュリティ、サイバーセキュリティ、内部脅威に対する安全対策を含む、強固なセキュリティ管理に投資し、実施する

[45] 広島 AI プロセスに関する G7 首脳声明「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」（2023 年 10 月）
なお、同文書は高度な AI システムにおける動向に対応して、既存の OECD AI 原則にもとづいて構築される living document であることに注意が必要である。
https://www.mofa.go.jp/mofaj/files/100573472.pdf

- 高度な AI システムのサイバーセキュリティリスクの評価や、適切で安全な環境での作業と文書保管の義務付けをすべき。無許可で公開されるリスク等に対応するための対策、知的財産や企業秘密の保護と整合性のある強固な内部脅威検知プログラムの確立すべき

VII. 技術的に可能な場合は、電子透かしやその他の技術等、AI 利用者及び業務外利用者が、AI が生成したコンテンツを識別できるようにするための、信頼できるコンテンツ認証及び来歴のメカニズムを開発し、導入する
- 透かしや識別子を利用することに加え、各主体がこの分野の状況を前進させるために協力し、研究に投資すべき

VIII. 社会的、安全、セキュリティ上のリスクを軽減するための研究を優先し、効果的な軽減策への投資を優先する
- 民主的価値の維持や人権の尊重、子どもや社会的弱者の保護等、リスクに対処するための優先的な研究、協力等を行うべき。加えて、環境や気候への影響を含むリスクを積極的に管理し、リスクに関する研究とベストプラクティスを共有することを奨励

IX. 世界の最大の課題、特に気候危機、世界保健、教育等（ただしこれらに限定されない）に対処するため、高度な AI システムの開発を優先する
- 個人や地域社会が AI の利用から利益を得るためのデジタル・リテラシーのイニシアティブを支援し、一般市民の教育と訓練を促進すべき。また、市民社会やコミュニティ・グループとの協力により、課題の特定や解決策を開発すべき

X. 国際的な技術規格の開発を推進し、適切な場合にはその採用を推進する
- 国際的な技術標準の開発に加え、AI が生成したコンテンツと他のコンテンツを区別できる技術標準を開発すべき

XI. 適切なデータインプット対策を実施し、個人データ及び知的財産を保護する
- データの質の管理のための適切な対策として、透明性、プライバシー保護のための機械学習や、機密データ等の漏洩のテストとファインチューニングを含む対策を実施し、著作権で保護されたコンテンツを含め、プライバシーや知的財産等に関する権利を尊重するために適切なセーフガードの導入が奨励される

当該「行動規範」の遵守状況を高度な AI システムを開発する AI 開発者自らが自主的に確認し報告する「報告枠組み」が OECD との協力の下、G7 で合意され、2025 年 2 月より運用開始されている[46]。当該「行動規範」を遵守した高度な AI システムを開発する AI 開発者は、「報告枠組み」に参加することが期待されている。

[46] 広島プロセス国際行動規範「報告枠組み」 https://transparency.oecd.ai/

第４部 AI 提供者に関する事項

AI 提供者は、AI 開発者が開発する AI システムに付加価値を加えて AI システム・サービスを AI 利用者に提供する役割を担う。AI を社会に普及・発展させるとともに、社会経済の成長にも大きく寄与する一方で、社会に与える影響の大きさゆえに、AI 提供者は、AI の適正な利用を前提とした AI システム・サービスの提供を実現することが重要となる。そのため、AI システム・サービスに組み込む AI が当該システム・サービスに相応しいものか留意することに加え、ビジネス戦略又は社会環境の変化によって AI に対する期待値が変わることも考慮して、適切な変更管理、構成管理及びサービスの維持を行うことが重要である。

AI システム・サービスを AI 開発者が意図している範囲で実装し、正常稼働及び適正な運用を継続し、AI 開発者に対しては、AI システムが適正に開発されるように求めることが重要である。AI 利用者に対しては、AI システムの提供及び運用のサポート又は AI システムの運用をしつつ AI サービスを提供することが重要である。提供に際し、ステークホルダーの権利を侵害せず、かつ社会に不利益等を生じさせることがないように留意し、合理的な範囲でインシデント事例等を含む関連情報の共有を行い、より安全安心で信頼できる AI システム・サービスを提供することが期待される。

以下に AI 提供者にとって、重要な事項を挙げる。

- **AI システム実装時**

  - P-2）i. 人間の生命・身体・財産、精神及び環境に配慮したリスク対策
    - AI 利用者を含む関連するステークホルダーの生命・身体・財産、精神及び環境に危害を及ぼすことがないよう、提供時点で予想される利用条件下でのパフォーマンスだけでなく、様々な状況下で AI システムがパフォーマンスレベルを維持できるようにし、リスク（連動するロボットの制御不能、不適切な出力等）を最小限に抑える方法（ガードレール技術等）を検討する（「2）安全性」）

  - P-2）ii. 適正利用に資する提供
    - AI システム・サービスの利用上の留意点を正しく定める（「2）安全性」）
    - AI 開発者が設定した範囲で AI を活用する（「2）安全性」）
    - 提供時点で AI システム・サービスの正確性・必要な場合には学習データの最新性（データが適切であること）等を担保する（「2）安全性」）
    - AI 開発者が設定した AI の想定利用環境と AI システム・サービスの利用者の利用環境に違い等がないかを検討する（「2）安全性」）

  - P-3）i. AI システム・サービスの構成及びデータに含まれるバイアスへの配慮
    - 提供時点でデータの公平性の担保及び参照する情報、連携する外部サービス等のバイアスを検討する（「3）公平性」）
    - AI モデルの入出力及び判断根拠を定期的に評価し、バイアスの発生をモニタリングする。また、必要に応じて、AI 開発者に AI モデルを構成する各技術要素のバイアスの再評価、評価結果

にもとづく AI モデル改善の判断を促す（「3）公平性」）

- ✧ AI モデルの出力結果を受け取る AI システム・サービス、ユーザーインタフェースにおいて、ビジネスプロセス及び AI 利用者又は業務外利用者の判断を恣意的に制限するようなバイアスが含まれてしまう可能性を検討する（「3）公平性」）

- ➢ P-4）i. プライバシー保護のための仕組み及び対策の導入
  - ✧ AI システムの実装の過程を通じて、採用する技術の特性に照らし適切に個人情報へのアクセスを管理・制限する仕組みの導入等のプライバシー保護のための対策を講ずる（プライバシー・バイ・デザイン）（「4）プライバシー保護」）

- ➢ P-5）i. セキュリティ対策のための仕組みの導入
  - ✧ AI システム・サービスの提供の過程を通じて、採用する技術の特性に照らし適切にセキュリティ対策を講ずる（セキュリティ・バイ・デザイン）（「5）セキュリティ確保」）

- ➢ P-6）i. システムアーキテクチャ等の文書化
  - ✧ トレーサビリティ及び透明性の向上のため、意思決定に影響を与える提供する AI システム・サービスのシステムアーキテクチャ、データの処理プロセス等について文書化する（「6）透明性」）

- **AI システム・サービス提供後**

  - ➢ P-2）ii. 適正利用に資する提供
    - ✧ 適切な目的で AI システム・サービスが利用されているかを定期的に検証する（「2）安全性」）

  - ➢ P-4）ii. プライバシー侵害への対策
    - ✧ AI システム・サービスにおけるプライバシー侵害に関して適宜情報収集し、侵害を認識した場合等は適切に対処するとともに、再発の防止を検討する（「4）プライバシー保護」）

  - ➢ P-5）ii. 脆弱性への対応
    - ✧ AI システム・サービスに対する攻撃手法も数多く生まれているため、最新のリスク及びそれに対応するために提供の各工程で気を付けるべき点の動向を確認する。また、脆弱性に対応することを検討する（「5）セキュリティ確保」）

  - ➢ P-6）ii. 関連するステークホルダーへの情報提供
    - ✧ 提供する AI システム・サービスについて、例えば以下の事項を平易かつアクセスしやすい形で、適時かつ適切に情報を提供する（「6）透明性」）
      - AI を利用しているという事実、活用している範囲、適切/不適切な使用方法等（「6）透明性」）
      - 提供する AI システム・サービスの技術的特性、利用によりもたらす結果より生じる可能性のある予見可能なリスク及びその緩和策等の安全性に関する情報（「2）安全性」）

- AI システム・サービスの学習等による出力又はプログラムの変化の可能性（「1）人間中心」）
- AI システム・サービスの動作状況に関する情報、不具合の原因及び対応状況、インシデント事例等（「2）安全性」）
- AI システムの更新を行った場合の更新内容及びその理由の情報（「2）安全性」）
- AI モデルにて学習するデータの収集ポリシー、学習方法、実施体制等（「3）公平性」、「4）プライバシー保護」、「5）セキュリティ確保」）

➢ P-7）i. AI 利用者への「共通の指針」の対応状況の説明

✧ AI 利用者に適正利用を促し、以下の情報を AI 利用者に提供する（「7）アカウンタビリティ」）

- 正確性・必要な場合には最新性（データが適切であること）等が担保されたデータの利用についての注意喚起（「2）安全性」）
- コンテキスト内学習による不適切な AI モデルの学習に対する注意喚起（「2）安全性」）
- 個人情報を入力する際の留意点（「4）プライバシー保護」）

✧ 提供する AI システム・サービスへの個人情報の不適切入力について注意喚起する（「4）プライバシー保護」）

P-7）ii. サービス規約等の文書化

✧ AI 利用者又は業務外利用者に向けたサービス規約を作成する（「7）アカウンタビリティ」）

✧ プライバシーポリシーを明示する（「7）アカウンタビリティ」）

なお、高度な AI システムを取り扱う AI 提供者は、「第 2 部 D. 高度な AI システムに関係する事業者に共通の指針」について以下のように対応する。

- I）～XI） 適切な範囲で遵守すべきである
- XII） 遵守すべきである

第５部 AI 利用者に関する事項

AI 利用者は、AI 提供者から安全安心で信頼できる AI システム・サービスの提供を受け、AI 提供者が意図した範囲内で継続的に適正利用及び必要に応じて AI システムの運用を行うことが重要である。これにより業務効率化、生産性・創造性の向上等 AI によるイノベーションの最大の恩恵を受けることが可能となる。また、人間の判断を介在させることにより、人間の尊厳及び自律を守りながら予期せぬ事故を防ぐことも可能となる。

AI 利用者は、社会又はステークホルダーから AI の能力又は出力結果に関して説明を求められた場合、AI 提供者等のサポートを得てその要望に応え理解を得ることが期待され、より効果的な AI 利用のために必要な知見習得も期待される。

以下に AI 利用者にとって、重要な事項を挙げる。

- **AI システム・サービス利用時**

  - U-2）i. 安全を考慮した適正利用
    - ✧ AI 提供者が定めた利用上の留意点を遵守して、AI 提供者が設計において想定した範囲内で AI システム・サービスを利用する（「2） 安全性」）
    - ✧ 正確性・必要な場合には最新性（データが適切であること）等が担保されたデータの入力を行う（「2） 安全性」）
    - ✧ AI の出力について精度及びリスクの程度を理解し、様々なリスク要因を確認した上で利用する（「2） 安全性」）

  - U-3）i. 入力データ又はプロンプトに含まれるバイアスへの配慮
    - ✧ 著しく公平性を欠くことがないよう公平性が担保されたデータの入力を行い、プロンプトに含まれるバイアスに留意して、責任をもって AI 出力結果の事業利用判断を行う（「3） 公平性」）

  - U-4）i. 個人情報の不適切入力及びプライバシー侵害への対策
    - ✧ AI システム・サービスへ個人情報を不適切に入力することがないよう注意を払う（「4） プライバシー保護」）
    - ✧ AI システム・サービスにおけるプライバシー侵害に関して適宜情報収集し、防止を検討する（「4） プライバシー保護」）

  - U-5）i. セキュリティ対策の実施
    - ✧ AI 提供者によるセキュリティ上の留意点を遵守する（「5） セキュリティ確保」）
    - ✧ AI システム・サービスに機密情報等を不適切に入力することがないよう注意を払う（「5） セキュリティ確保」）

  - U-6）i. 関連するステークホルダーへの情報提供

- ✧ 著しく公平性を欠くことがないよう、公平性が担保されたデータの入力を行い、プロンプトに含まれるバイアスに留意して AI システム・サービスから出力結果を取得する。そして、出力結果を事業判断に活用した際は、その結果を関連するステークホルダーに合理的な範囲で情報を提供する（「3）公平性」、「6）透明性」）

➢ U-7）i.関連するステークホルダーへの説明

- ✧ 関連するステークホルダーの性質に応じて合理的な範囲で、適正な利用方法を含む情報提供を平易かつアクセスしやすい形で行う（「7）アカウンタビリティ」）
- ✧ 関連するステークホルダーから提供されるデータを用いることが予定されている場合には、AI の特性及び用途、データの提供元となる関連するステークホルダーとの接点、プライバシーポリシー等を踏まえ、データ提供の手段、形式等について、あらかじめ当該ステークホルダーに情報提供する（「7）アカウンタビリティ」）
- ✧ 当該 AI の出力結果を特定の個人又は集団に対する評価の参考にする場合は、AI を利用している旨を評価対象となっている当該特定の個人又は集団に対して通知し、当ガイドラインが推奨する出力結果の正確性、公正さ、透明性等を担保するための諸手続きを遵守し、かつ自動化バイアスも鑑みて人間による合理的な判断のもと、評価の対象となった個人又は集団からの求めに応じて説明責任を果たす（「１）人間中心」、「6）透明性」、「7）アカウンタビリティ」）
- ✧ 利用する AI システム・サービスの性質に応じて、関連するステークホルダーからの問合せに対応する窓口を合理的な範囲で設置し、AI 提供者とも連携の上説明及び要望の受付を行う（「7）アカウンタビリティ」）

➢ U-7）ii.提供された文書の活用及び規約の遵守

- ✧ AI 提供者から提供された AI システム・サービスについての文書を適切に保管・活用する（「7）アカウンタビリティ」）
- ✧ AI 提供者が定めたサービス規約を遵守する（「7）アカウンタビリティ」）

なお、高度な AI システムを取り扱う AI 利用者は、「第 2 部 D. 高度な AI システムに関係する事業者に共通の指針」について以下のように対応する。

- I）～XI） 適切な範囲で遵守すべきである
- XII） 遵守すべきである