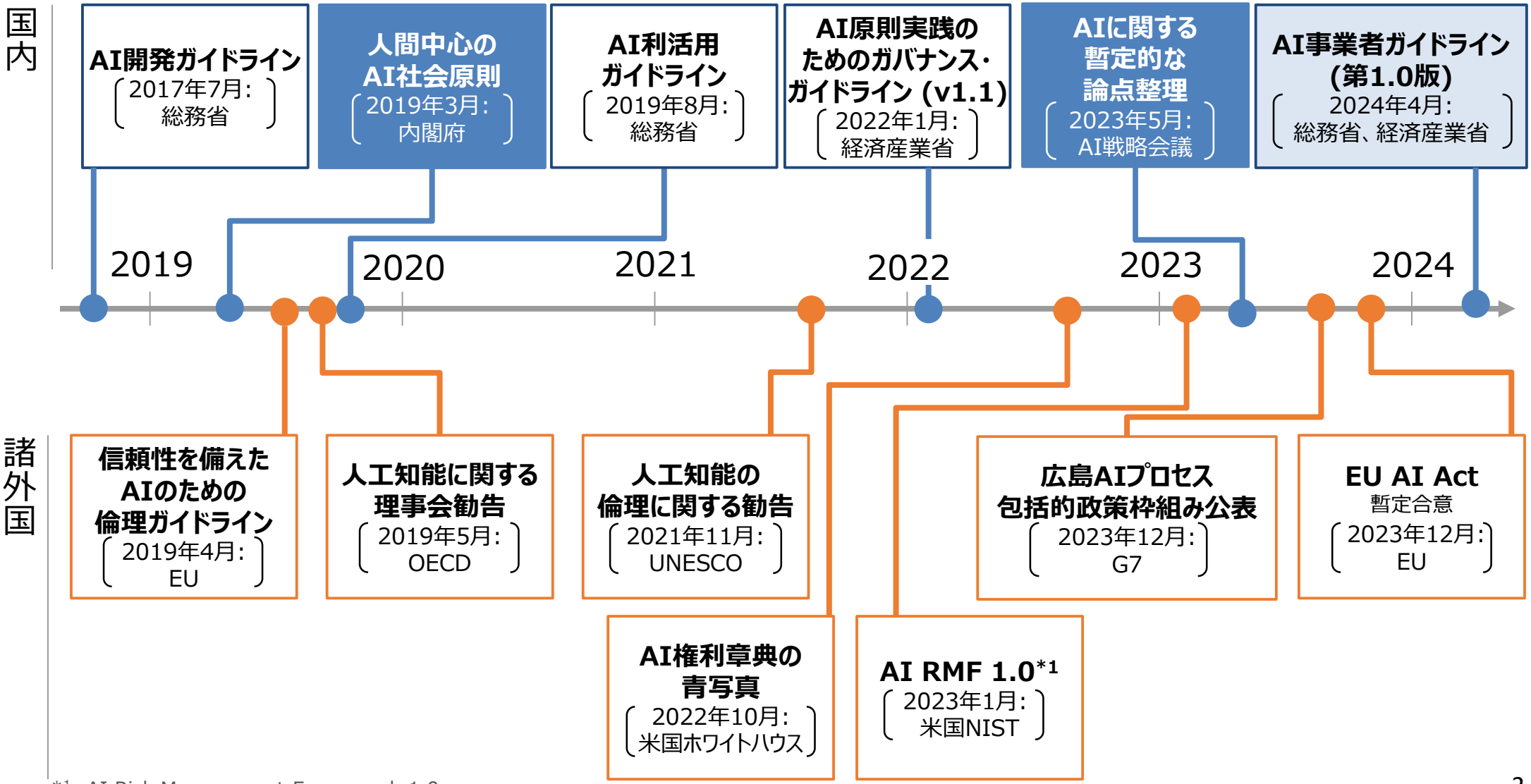
AI事業者ガイドライン(第1.0版)について

経済産業省　商務情報政策局　情報経済課
ガバナンス戦略国際調整官　飯野　悠介
令和６年８月２３日

【参考】ＡＩ事業者ガイドラインの経緯

- 「人間中心のAI社会原則」において、AIは、「社会に多大なる便益をもたらす一方で、その社会への影響が大きいゆえに、適切な開発と社会実装が求められる」ことが指摘。
- 「AIに関する暫定的な論点整理」において、AIの開発・提供・利用を促進しつつも、リスクへの適切な対処「ガードレール」が必要であると言及。



*1: AI Risk Management Framework 1.0

「AI事業者ガイドライン」の基本的な考え方

- 本ガイドラインは、「1 事業者の自主的な取組の支援」、「2 国際的な議論との協調」、「3 読み手にとっての分かりやすさ」を基本的な考え方としている
- 加えて、「マルチステークホルダー」で検討を重ね実効性・正当性を重視するとともに、「Living Document」として今後も更新を重ねていく

考え方

1 事業者の自主的な取組の支援

対策の程度をリスクの大きさ及び蓋然性に対応させる「リスクベースアプローチ」に基づく企業における対策の方向性を記載

2 国際的な議論との協調

国内外の関連する諸原則の動向や内容との整合性を確保

3 読み手にとっての分かりやすさ

「AI開発者」・「AI提供者」・「AI利用者」ごとに、AIに関わる考慮すべきリスクや対応方針を確認可能

＋

プロセス

マルチステークホルダー

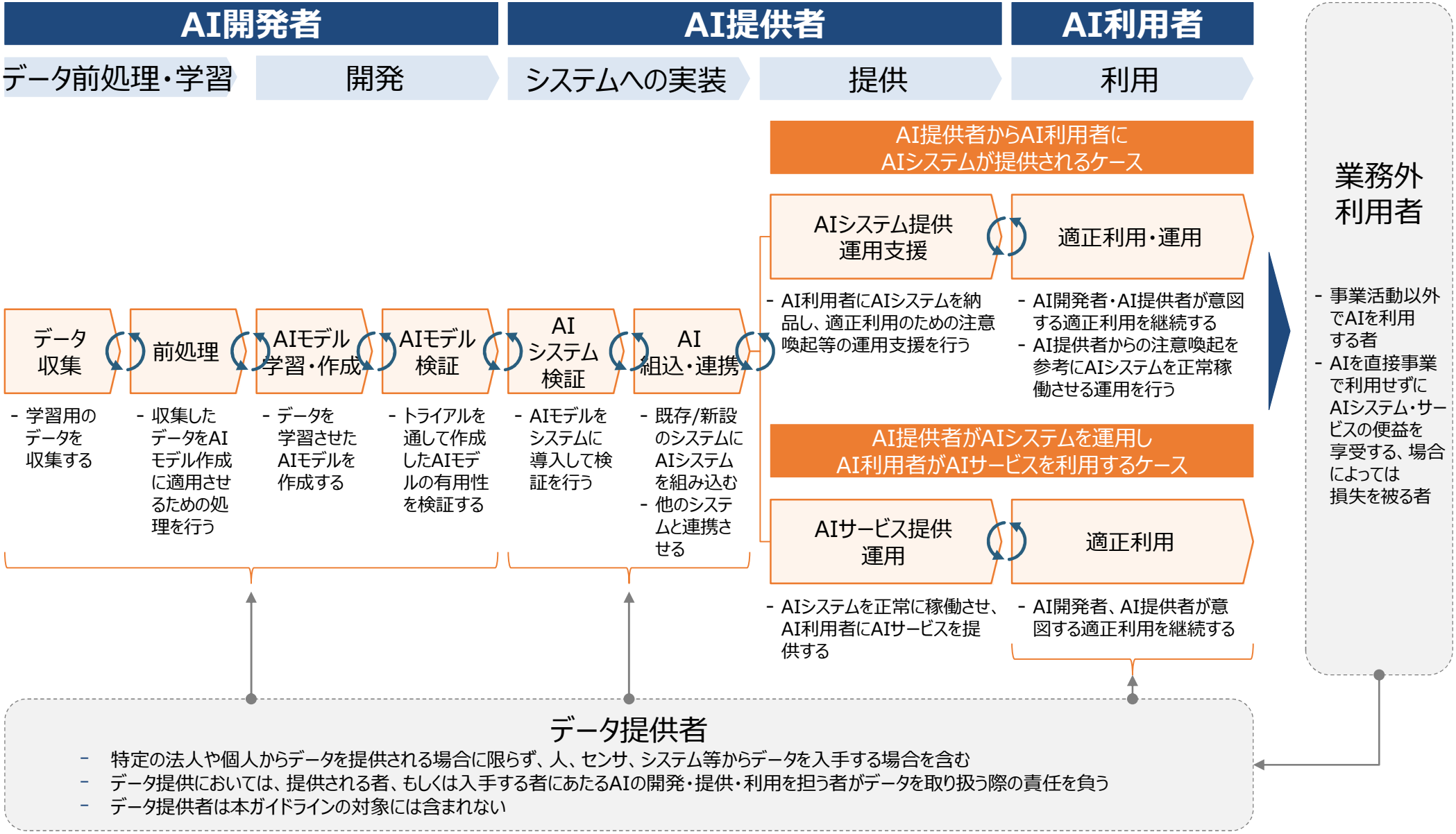
教育・研究機関、一般消費者を含む市民社会、民間企業等で構成されるマルチステークホルダーで検討を重ねることで、実効性・正当性を重視したものとして策定

Living Document

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適宜、更新

一般的なAIの事業活動を担う主体

- AIライフサイクルにおける具体的な役割を考慮し、AIの事業活動を担う立場として、「AI開発者」、「AI提供者」、「AI利用者」の3つに大別して整理する　※「データ提供者」、「業務外利用者」は対象外とする



「AI事業者ガイドライン」本編、別添の位置づけ

- 本編では、事業者がAIの安全安心な活用を行い、AIの便益を最大化するために重要な「どのような社会を目指すのか（基本理念＝why）」及び「どのような取組を行うか（指針＝what）」を示した
- 別添（付属資料）では、「具体的にどのようなアプローチで取り組むか（実践＝how）」を示すことで、事業者の具体的な行動へとつなげることを想定している

本編（why, what） ▶ **別添（付属資料）(how)**

どのような社会を
目指すのか
（基本理念＝why）

どのような取組を
行うか
（指針＝what）

どのようなアプローチで
取り組むか
（実践＝how）

【参考】別添1 A. AIに関する前提　主な記載内容

- 本編での各主体・定義等を具体化するAIそのもの、活用場面及び各主体の役割の更なる理解につなげる
  - 主なAIシステム・サービスを例示しつつ、関連する主体を具体化する
  - ユースケース毎に、プロセスと各主体 (AI事業者) の役割及びパターンを整理する

AIシステム・サービスの例　(抜粋)

ケース名	活用 AI	概要	AI 開発者	AI 提供者	AI 利用者	業務外利用者
採用 AI	テキスト解析	A 社グループのグローバル各社における人材採用部門が、**エントリーシートの書類選考を判断する際の参考情報として使用される AI サービス**である。 A 社 AI 開発部門は、AI 利用者である A 社人材採用部門（海外グループ企業を含む）より過去のエントリーシートデータ及び合否判定（内定の判定）結果を受領し、機械学習（分類モデル）で合否判定を支援する AI モデルを作成している。	A 社（開発部門）	A 社（システム部門、人材開発部門）	A 社グループ（人材採用部門）	採用申込者
無人コンビニ	画像解析	全国のコンビニエンスストアチェーンを経営する J 社が提供する**画像認識 AI を活用した無人コンビニ（店内の客が商品を取るだけで AI が代金を計算し、店外に出る際に電子マネー等で一括決済ができるコンビニ）**である。当 AI サービスには X 社で開発された無人コンビニ向けの AI システムを搭載している。	X 社	J 社（AI システム開発部及びコンビニ事業部）	コンビニ店舗	コンビニ利用客
がん診断 AI	テキスト・画像解析	マルチモーダル学習を使用しており、**「本人の病歴・遺伝等に係る情報（データ 1）」及び「内視鏡画像（データ 2）」を取込み、内視鏡での診察中にリアルタイムにがんの可能性が高い部分をハイライト**する。医師は出力画像を観察して、がんの可能性があるか判断する。 A 社が AI を開発しつつ、がん診断 AI システムを医療機関に提供している。	A 社（AI 開発部門）	A 社（医療 IT サービス部門）	医療機関（システム部門及び消化器内科）	受診患者

AI事業者のパターン

【参考】別添1 B. AIによる便益/リスク（AIによる便益）主な記載内容

- 便益を享受する最終利用者に焦点を当ててAIによる便益を整理し、その理解促進につなげる

	開発	マーケティング	販売	物流・流通	顧客対応	法務	ファイナンス	人事
従来から存在する便益の例	コード検証、ドキュメント作成の自動化	広告用メールの自動配信	受注後の対応メール等の自動発信	需要予測に基づく生産・在庫数最適化	チャットボットによる自動対応	翻訳	財務諸表の自動作成	給与計算等の自動化
（生成AIで更に向上）	類似コード・データの抽出・検証	データに基づいたパーソナライゼーション広告	チャネル別、ニーズ別の売上予測	配送ルート最適化	過去の問合せ内容に基づいたFAQ作成	法務文章のレビュー	過去実績にもとづいた将来予測、不正検知	職務経歴書等に基づいた人材需要マッチング
生成AI特有の便益の例	学習データの生成、コーディングアシスタント、新製品のブレインストーミング	販売促進(マーケティング素材・キャッチコピー等)の自動作成	営業トークスクリプトの自動作成	物流条件交渉のアシスタント	対応内容の自動生成、要約	規定に基づいた契約書ドラフトの自動生成	文脈を踏まえた上での社内問合せ対応	文脈を踏まえた上での人事面接の対応

【参考】別添1 B. AIによる便益/リスク（AIによるリスク）主な記載内容

- 従来型のAIからのリスク及び生成AIで特に顕在化したリスクについて、主なものを整理
- AIによるリスクを知ることで、AIガバナンスの必要性の理解を深める

	リスク	事例	対応する「共通の指針」
従来型AIから存在するリスク	バイアスのある結果及び差別的な結果の出力	• IT企業が自社で開発したAI人材採用システムが女性を差別するという機械学習面の欠陥を持ち合わせていた	1) 人間中心 3) 公平性
	フィルターバブル及びエコーチェンバー現象	• SNS等によるレコメンドを通じた社会の分断が生じている	1) 人間中心
	多様性の喪失	• 社会全体が同じモデルを、同じ温度感で使った場合、導かれる意見及び回答がLLMによって収束してしまい、多様性が失われる可能性がある	1) 人間中心
	不適切な個人情報の取扱い	• 透明性を欠く個人情報の利用及び個人情報の政治利用も問題視されている	1) 人間中心 4) プライバシー保護
	生命、身体、財産の侵害	• AIが不適切な判断を下すことで、自動運転車が事故を引き起こし、生命や財産に深刻な損害を与える可能性がある • トリアージにおいては、AIが順位を決定する際に倫理的なバイアスを持つことで、公平性の喪失等が生じる可能性がある	2) 安全性 3) 公平性
	データ汚染攻撃	• AIの学習実施時及びサービス運用時には学習データへの不正データ混入、サービス運用時ではアプリケーション自体を狙ったサイバー攻撃等のリスクが存在する	5) セキュリティ確保
	ブラックボックス化、判断に関する説明の要求	• AIの判断のブラックボックス化に起因する問題も生じている • AIの判断に関する透明性を求める動きも上がっている	6) 透明性 7) アカウンタビリティ
	エネルギー使用量及び環境の負荷	• AIの利用拡大により、計算リソースの需要も拡大しており、結果として、データセンターが増大しエネルギー使用量の増加が懸念されている	1) 人間中心
生成AIで特に顕在化したリスク	悪用	• AIの詐欺目的での利用も問題視されている	2) 安全性 8) 教育・リテラシー
	機密情報の流出	• AIの利用においては、個人情報や機密情報がプロンプトとして入力され、そのAIからの出力等を通じて流出してしまうリスクがある	5) セキュリティ確保 8) 教育・リテラシー
	ハルシネーション	• 生成AIが事実と異なることをもっともらしく回答する「ハルシネーション」に関してはAI開発者・提供者への訴訟も起きている	2) 安全性 8) 教育・リテラシー
	偽情報、誤情報を鵜呑みにすること	• 生成AIが生み出す誤情報を鵜呑みにすることがリスクとなりうる • ディープフェイクは、各国で悪用例が相次いでいる	1) 人間中心 8) 教育・リテラシー
	著作権との関係	• 知的財産権の取扱いへの議論が提起されている	2) 安全性
	資格等との関係	• 生成AIの活用を通じた業法免許や資格等の侵害リスクも考えうる	2) 安全性
	バイアスの再生成	• 生成AIは既存の情報に基づいて回答を作るため既存の情報に含まれる偏見を増幅し、不公平や差別的な出力が継続/拡大する可能性がある	3) 公平性

基本理念

- 「人間中心のAI社会原則」において、AIがSociety 5.0の実現に貢献し、AIを人類の公共財として活用することで、社会の在り方の質的変化や真のイノベーションを通じて地球規模の持続可能性へとつなげることが重要であると述べられている
- 加えて、下記の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」としており、この普遍的な考え方は、今後も目指すべき理念であり続けている

人間の尊厳が尊重される社会
(Dignity)

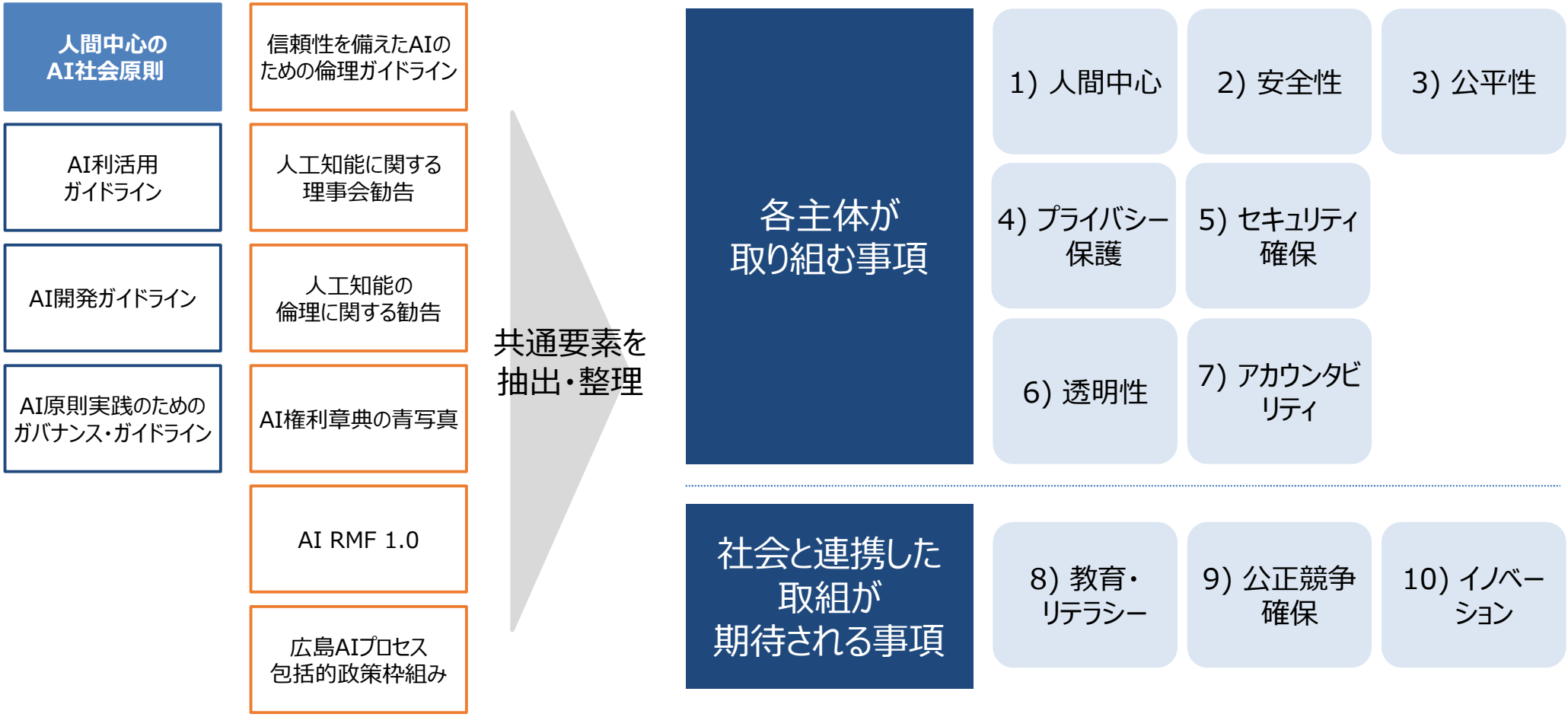
基本理念

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 & Inclusion)

持続可能な社会
(Sustainability)

各主体に共通の指針

- AIの活用による目指すべき社会の実現のために各主体が連携して取り組む内容を原則としてまとめた上で、「共通の指針」として整理する
- 「共通の指針」は、「人間中心のAI社会原則」を土台としつつ、諸外国における議論状況や、新技術の台頭に伴い生じるリスクへの対応等を反映している
- その結果、各主体が取り組む事項、及び社会と連携して取り組むことが期待される事項に分類される



【参考】各主体に共通の指針 [1/2]

- 各主体は、1）人間中心に照らし、法の支配、人権、民主主義、多様性及び公平公正な社会を尊重する
- 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
- AIガバナンスを構築し継続的に運用（AIのもたらすリスクの程度や各主体の資源制約に配慮しつつ実施）

指針		内容（主な項目の抜粋）
各主体が取り組む事項	1) 人間中心	✓ 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 ✓ 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 ✓ 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
	2) 安全性	✓ 適切なリスク分析を実施し、**リスクへの対策**を講じる ✓ 主体のコントロールが及ぶ範囲で本来の利用目的を逸脱した提供・利用により危害が発生することを避ける ✓ 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等を検討するとともに、**データの透明性の支援、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
	3) 公平性	✓ 特定の個人ないし集団へのその人種、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 ✓ 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した上で、無意識や潜在的な**バイアスに留意**し、AIの開発・提供・利用を行う
	4) プライバシー保護	✓ 個人情報保護法等の**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**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
	5) セキュリティ確保	✓ 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

【参考】各主体に共通の指針 [2/2]

- 各主体は、1）人間中心に照らし、法の支配、人権、民主主義、多様性及び公平公正な社会を尊重する
- 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
- AIガバナンスを構築し継続的に運用（AIのもたらすリスクの程度や各主体の資源制約に配慮しつつ実施）

指針		内容（主な項目の抜粋）
各主体が取り組む事項（続き）	6) 透明性	✓ AI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適切な情報を提供**する（AIを利用しているという事実、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、AIシステム・サービスの能力、限界、提供先における適切/不適切な利用方法、等）
	7) アカウンタビリティ	✓ トレーサビリティの確保や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情報の提供と説明を行う ✓ 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 ✓ 関係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
社会と連携した取組が期待される事項	8) 教育・リテラシー	✓ 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 ✓ 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。
	9) 公正競争確保	✓ 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が維持**に努めることが期待される
	10) イノベーション	✓ 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 ✓ 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

【参考】高度なAIシステムに関係する事業者に共通の指針

- 「共通の指針」に加え、以下を遵守すべきである[*1]。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、各主体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

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

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

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

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基づく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

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

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等が、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

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

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）に対処する対処するため、高度なAIシステムの開発を優先する

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

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

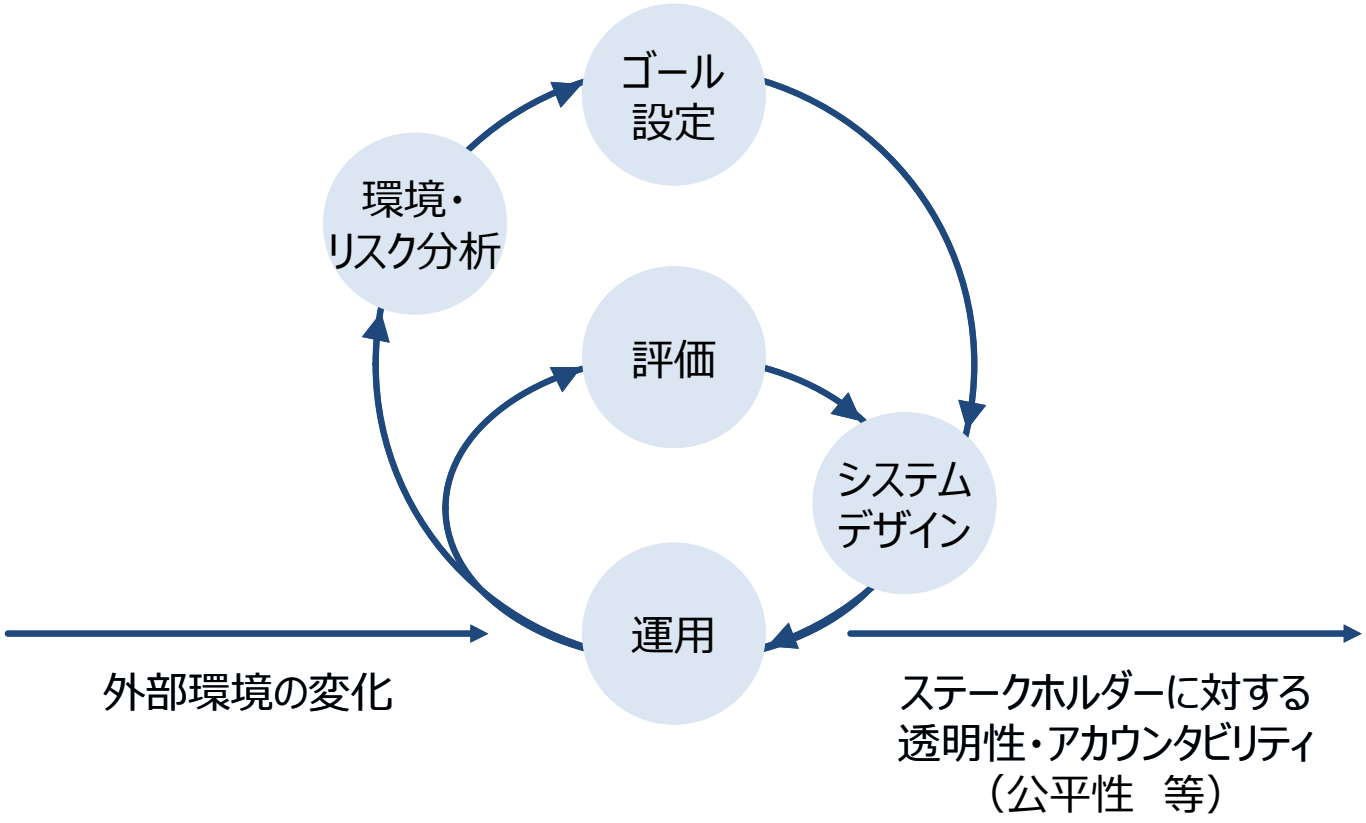
XII. 高度な AI システムの信頼でき責任ある利用を促進し、貢献する

*1: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。

AIガバナンスの構築

- AIを安全安心に活用していくために、経営層のリーダーシップのもと、下記に留意しながら適切なAIガバナンスを構築することで、リスクをマネジメントしていくことが重要となる
  - 複数主体に跨る論点について、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確保
  - 上記が複数国にわたる場合、データの自由な越境移転の確保のための適切なAIガバナンスの検討
  - 経営層のコミットメントによる、各組織の戦略や企業体制への落とし込み/文化としての浸透

適切なAIガバナンスの構築



複数主体間の連携確保

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

適切なデータ流通

複数国にまたがる想定の場合の適切なリスク管理/AIガバナンスの実施

経営層のコミットメント

戦略/体制への落とし込み、各組織の文化としての浸透

【参考】別添2. 第2部 [E.AIガバナンスの構築]関連

- 別添2では、AIガバナンスの構築のための「行動目標」、「実践のポイント」及びそれに対応する仮想的な「実践例」、実際の企業の取組事例を掲載
- 各事業者にとって、自身のAIガバナンスの構築のための検討を、具体例を交えつつ行うことを可能とする

A.経営層によるAIガバナンスの構築及びモニタリング

- 行動目標
  - 一般的かつ客観的な目標を記載
- 実践のポイント
  - 「AI原則実践のためのガバナンスガイドライン Ver. 1.1」を基に、国内外ガイドラインやISO等の要素も取り入れ
- 実践例
  - 仮想事例に基づく事例を記載
  - 生成AI等の最新動向への対応事例も織り込み

例）行動目標1-1【便益/リスクの理解】【実践例v：生成AI への対応】
「自社の業務に活用するにあたり、JDLA〔※〕の発行する「生成AI の利用ガイドライン」に沿って、自社内での利用ガイドラインの策定を行っている。併せて、個人情報保護委員会の「生成 AI サービスの利用に関する注意喚起等」等の行政からの情報発信も確認している。また、ニュース及びSNS 等を通じ生成AI に関する情報の収集も欠かせない。これらを通じて、便益及びリスクの最新動向の把握に努めている。」〔※一般社団法人日本ディープラーニング協会〕

B.AIガバナンスの構築に関する実際の取組事例

- 実際の取組事例
  - 「AI原則実践のためのガバナンスガイドライン Ver. 1.1」に基づくAIガバナンスの取組事例について、5社分をコラム化
  - AIガバナンスを推進するにあたり、多くの企業がつまづく観点を記載

【参考】別添3～5．各主体向け構成

- 各主体の重要事項に対して、「ポイント」、「具体的な手法」、「参考資料」を以下の構成で掲載

本編の
第3~5部に
対応する解説

本編 第2部に
対応する解説
「共通の指針」に
関する各主体の
具体的なアプローチ

別添　記載内容

A. 本編「第 3 部　AI 開発者に関する事項」の解説

[本編の記載内容 （再掲）]
データ前処理・学習時
D-2） i. 適切なデータの学習

Article I.　　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」）
Article II.　　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」）

[ポイント]
AI モデルの質の向上のために、AI 開発者は、AI の学習等に用いるデータの質に留意することが重要となる。
- 利用する AI の特性及び用途を踏まえ、AI の学習等に用いるデータの質（正確性及び完全性等）に留意する
- また、AI によりなされる判断は、事後的に精度が損なわれたり、低下したりすることが想定されるため、想定される権利侵害の規模、権利侵害の生じる頻度、適用できる技術水準、精度を維持するためのコスト等を踏まえ、あらかじめ精度に関する基準を定めておくことが期待される。精度が当該基準を下回った場合には、データの質に留意して改めて学習させる
- なお、ここで言う「精度」には、AI が倫理的に正しい判断を行っているか（例えば、AI が暴力的な表現を行っていないか、ヘイトスピーチ等を行っていないか　等）も含まれる

[具体的な手法]
1. データに個人情報、機密情報、著作権等の権利又は法律上保護される利益に関係するものが含まれていないか、確認を実施
   1.1. 固有表現抽出

[参考文献]
1. 国立研究開発法人産業技術総合研究所「機械学習品質マネジメントガイドライン 第 4 版」（2023 年 12 月）
2. NIST, “AI Risk Management Framework Playbook”（2023 年 1 月）

解説

本編の記載内容を再掲

ポイント
- 本編記載事項に加え、重要観点を補足

具体的な手法
- 他のガイドライン等を参照しつつ具体的に解説

参考文献
- 「ポイント」や「具体的な手法」の参照元を記載

【参考】AI開発者に関する事項 [1/2]

- AI開発者は、AIモデルを直接的に設計・変更ができるため、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特に重要

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、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、法令に則って適切に扱う - 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
	D-3) i.	データに含まれるバイアス等への配慮	- 学習データ、モデルの学習過程でバイアス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 - 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
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	- 予期しない環境を含む様々な状況下での利用に耐えうる性能の要求 - **リスクを最小限に抑える**方法の要求
	D-2) iii.	適正利用に資する開発	- **AIを安全に利用可能な使い方について明確な方針・ガイダンスを設定**する - 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
	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**各技術要素によってバイアスが含まれうる**ことまで検討する - 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する
	D-5) i.	セキュリティ対策のための仕組みの導入	- 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）
	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

【参考】AI開発者に関する事項 [2/2]

- AI開発者は、AIモデルを直接的に設計・変更ができるため、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特に重要

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
D-6) ii.	関連するステークホルダーへの情報提供	- AIシステムの技術的特性、安全性確保の仕組み、予見可能なリスク及びその緩和策、不具合の原因及び対応状況等に関する**情報提供**を行う
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に活用開始後に品質が変動する可能性及び、その結果として**生じるリスク等の情報提供及び説明**を行う
D-7) ii.	開発関連情報の文書化	- 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**文書化**する
D-10) i.	イノベーションの機会創造への貢献	- 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解決**につながるよう貢献する - 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等の取組を行う等、国際化・多様化及び産学官連携を行う - **社会全体への情報提供**を行う

最先端の基盤モデル及び生成AIシステムを含む、高度なAIシステムを開発するAI開発者については「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」を参照

【参考】AI提供者に関する事項 [1/2]

- AI提供者は、AIの稼働と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

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	- 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、**リスクを最小限に抑える**方法を検討する
P-2) ii.	適正利用に資する提供	- AI開発者が設定した範囲でAIを活用する - AIシステム・サービスの正確性等を担保すると同時に、**AI開発者の想定利用環境とAI利用者の利用環境に違い等がないか検討**する
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	- データの公平性を担保し、参照する情報、外部サービス等の**バイアスを検討**する - 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 - AIモデルの出力結果を受け取るAIシステム等において、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る可能性を検討する
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	- 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**プライバシー保護対策を講ずる**（プライバシー・バイ・デザイン）
P-5) i.	セキュリティ対策のための仕組みの導入	- 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）
P-6) i.	システムアーキテクチャ等の文書化	- AIシステムの意思決定に影響を与える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

【参考】AI提供者に関する事項 [2/2]

- AI提供者は、AIの稼働と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

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
	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
	P-5) ii.	脆弱性への対応	- 最新のリスク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し、**脆弱性に対応することを検討する**
	P-6) ii.	関連するステークホルダーへの情報提供	- AIシステムの・サービスの技術的特性、予見可能なリスク、緩和策、出力又はプログラムの変化の可能性、不具合の原因と対応状況、インシデント事例、学習データの収集ポリシー、その学習方法等に関する情報を説明できるようにする - AIの性質及び利用目的等に照らして、**AIを利用しているという事実や適切/不適切な使用方法、更新内容とその理由等の情報提供や説明の実施**
	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**、正確性・必要に応じて最新性等が担保されたデータの利用やコンテキスト内学習による不適切なモデルの学習に対する注意喚起、**個人情報を入力する際の留意点についての情報を提供**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
	P-7) ii.	サービス規約等の文書化	- AI利用者に向けた**サービス規約を作成するとともにプライバシーポリシーを明示**する

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、
I）～XI）を適切な範囲で遵守し、XII)について遵守すべき

【参考】AI利用者に関する事項

- AI利用者は、AI提供者が意図した範囲内で継続的に適正利用、必要に応じたAIシステムの運用を行うことが重要であり、より効果的なAI利用のために必要な知見を習得することが期待される

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利用**する - 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
	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	- 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**
	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しないよう注意を払う - 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
	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 - AIシステム・サービスに機密情報等を不適切に入力しないよう注意を払う
	U-6) i.	関連するステークホルダーへの情報提供	- 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、**出力結果を取得し、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提供**する
	U-7) i.	関連するステークホルダーへの説明	- AIの特性や用途、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平易かつアクセスしやすい方法で情報提供**する - 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とする場合は、人間による合理的な判断のもと、説明責任を果たす - **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
	U-7) ii.	提供された文書の活用と規約の遵守	- 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**文書を保管・活用**する - AI提供者が定めた**サービス規約を遵守**する

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、I）～XI）を適切な範囲で遵守し、XII)について遵守すべき

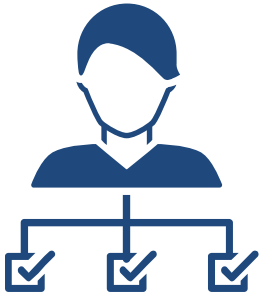
別添7. チェックリスト 概要

- 別添7では、AIによるリスクを抑えつつ便益を享受する取組の立案、実践を確実に推進するための「チェックリスト」及び「具体的なアプローチのためのワークシート」を用意している

本編・別添1～5（付属資料） » **チェックリスト** » **具体的なアプローチ検討のためのワークシート**

本編・別添を読んでAIガバナンスの重要性や、各事業者に期待されることを**理解する**

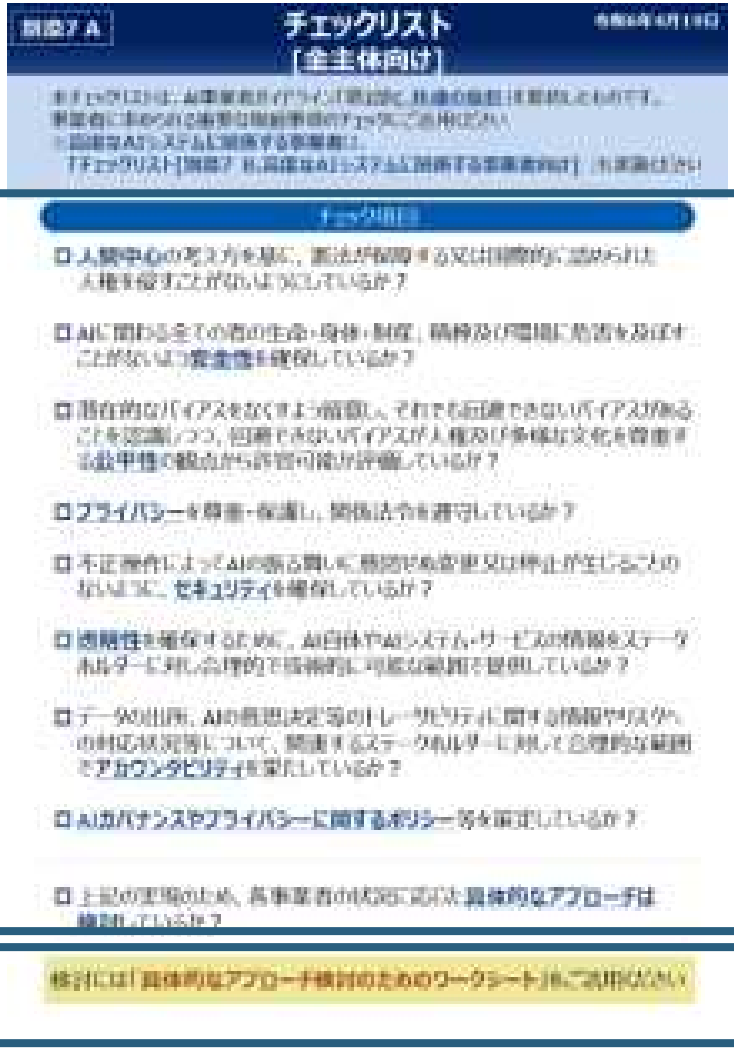
「チェックリスト」を活用し、本編・別添についての各主体の**取組(What)を確認**する



「具体的なアプローチ検討のためのワークシート」を使用し、各事業者の**具体的なアプローチ(How) を検討する**

別添7. チェックリスト 活用方法

- 全事業者は、「別添7 Aチェックリスト[全主体向け]」を活用し、各自の取組状況の概観を確認する
- 高度なAIシステムに関係する事業者に該当する場合には、「別添7 Bチェックリスト[高度なAIシステムに関係する事業者向け]」も実施する


別添7 A
チェックリスト
[全主体向け]

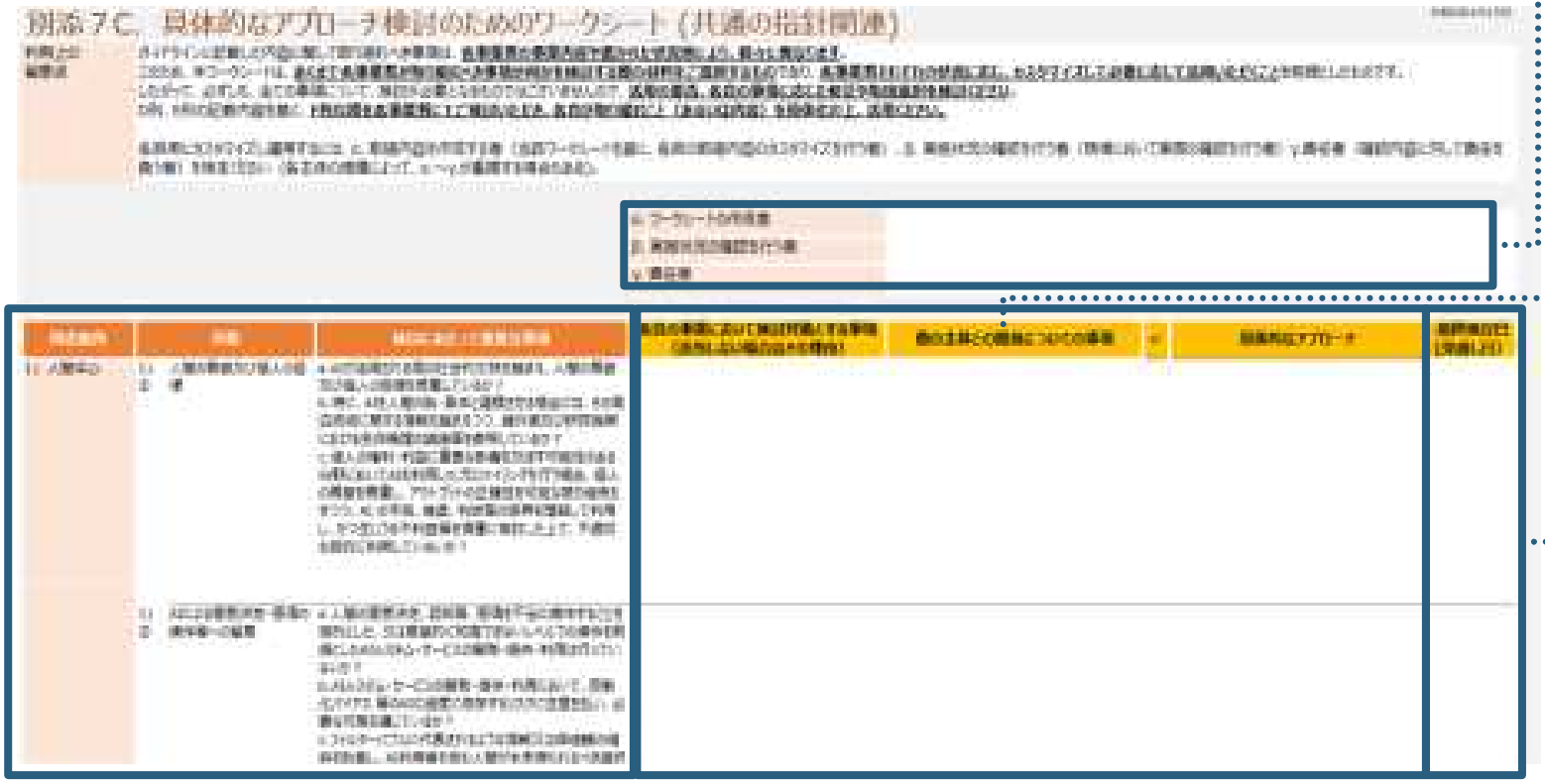

別添7 B
チェックリスト
[高度なAIシステムに関係する事業者向け]

- チェック項目は本編の要約を記載
- チェックすることで、**各自の取組状況を概観**

- 具体的な実践内容の検討に、**「別添7C 具体的なアプローチのためのワークシート」**を活用
(活用方法次頁)

別添7. 具体的なアプローチ検討のためのワークシート活用方法

- 本ガイドラインの記載内容に関して、具体的なアプローチを検討する際に重要となる事項を記載
- 事業者の事業内容及び置かれた状況等に応じ、各自でカスタマイズして活用することを前提としている



- 活用の前に、**実施責任者等を決定**する
- 事業者の事業内容や置かれた状況等に応じ、**各自でカスタマイズして活用**する
- 「見直し日」も検討・記載することで**定期的な更新**を行う
- 具体のアプローチを検討する際に重要になる事項が記載されており、事業者が**各自でカスタマイズする際のリファレンス**となる

【参考】別添8. 主体横断的な仮想事例 概要

- 別添8.では、本ガイドラインに沿って、AI開発者、AI提供者、AI利用者が重要事項の検討を行った場合の「主体横断的な仮想事例」を掲載
- 各主体が本ガイドラインの内容を実際に落とし込む際の具体的なイメージの想起や、各主体間での連携が重要になるポイントの明確化が可能
- 現在「採用AI」を扱う事業者を例として取り上げているが、今後、他の事例も追加していく予定

Case	採用AI			
機械学習モデル：XGBoost（エントリーシートの文章で、応募者に対して合否を判断する）				
https://ilt.u-tokyo.ac.jp/wp/wp-content/uploads/2022/10/ROWalet_CaseG1_Recruitment-AI_JP.pdf				
		AI開発部門	人材採用部門（採用AIチーム）	人材採用担当者
		AI開発者	AI提供者	AI利用者
No.	分類	共通の指針/各主体に関する原則	本AIにおいて主体が実施している活動	本AIにおいて主体が実施している活動	本AIにおいて主体が実施している活動
Ⅰ）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適切とされる主体として、少なくとも国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、多様な人々が多様な幸せを実現できる社会を目指すべきである。				
①人間の尊厳及び個人の自律				
1	共通	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	AIシステムの開発において、学習データの収集やラベリング、モデルの性能評価等は、AI開発者だけで完結せず、AI提供者等と連携して実施する。	AIサービスの提供において、AI利用者が最終判断（応募者の合否）を行えるようになっている（Human-in-the-loop）	AIサービスの提供において、AI利用者が最終判断（応募者の合否）を行えるようになっている（Human-in-the-loop）
2	共通	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	脳・身体と連携するケースではないため対象外	脳・身体と連携するケースではないため対象外	脳・身体と連携するケースではないため対象外
3	共通	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	AIシステムの開発において、実際の予測結果を学習データに用いる際には個人情報の取扱いに関する契約書の締結やアクセス権管理等を実施している。 ※公平性とプライバシーについては、「3）公平性」「4）プライバシー保護」を参照	AIシステムの開発において、実際の予測結果を学習データに用いる際には個人情報の取扱いに関する契約書の締結やアクセス権管理等を実施している。 ※公平性とプライバシーについては、「3）公平性」「4）プライバシー保護」を参照	AIシステムの開発において、実際の予測結果を学習データに用いる際には個人情報の取扱いに関する契約書の締結やアクセス権管理等を実施している。 ※公平性とプライバシーについては、「3）公平性」「4）プライバシー保護」を参照
②AIによる意思決定・感情の操作等への留意				
1	共通	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	本ケースに関しては、2)②-3と同じ論点になる	本ケースに関しては、2)②-3と同じ論点になる	本ケースに関しては、2)②-3と同じ論点になる
2	共通	AIシステムの開発・提供・利用において、自動化バイアス等の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	本ケースに関しては、2)②-3と同じ論点になる	本ケースに関しては、2)②-3と同じ論点になる	本ケースに関しては、2)②-3と同じ論点になる

ご清聴ありがとうございました