

第9回 AI倫理・AIガバナンスと医療の邂逅

(アンダーソン・毛利・友常法律事務所外国法共同事業 弁護士 中崎 尚)

1. AI倫理とは

AI倫理について、生成AIが中心に議論をされることが多いが、AIガバナンスに関しては生成AIに限らず、AI全体を幅広くに議論されることが多いです。

ANDERSON MÖRI & TOMOTSUNE

I-1. AI倫理

<AI倫理とは>

- AI倫理とは、「明確な定義は存在しないが、自律的」「な意思決定を行うAIが有すべき倫理の内容やAIに倫理を持たせるための手段に関する問題領域を指す」と説明されている。
- 国内外で数多くのガイドラインが策定・公表されているが、その内容にはズレが見られる。

AI倫理は、医療関係や生命倫理は異なる独自の概念です。明確な定義はないが、「AIが有すべき倫理的な内容やAIに倫理を持たせるための手段に関する領域を指す」と説明されています。

AI倫理に基づいて、例えばヨーロッパの「trustworthy」のガイドラインや日本国内で総務省が発表したAI倫理のガイドラインなど、様々な国で策定しています。

I - 1. AI倫理

<国内のAI倫理関連のガイドライン・原則>

- 「国際的な議論のための AI 開発ガイドライン案」（通称：AI 開発ガイドライン）（総務省 AI ネットワーク社会推進会議 2017年）
- 「人間中心の AI 社会原則」（内閣府 2019年）
- 「AI 利活用ガイドライン」（総務省 AI ネットワーク社会推進会議 2019年）
- 「AI 事業者ガイドライン（第1.0版）」（総務省・経済産業省 2024年4月）

2017年に総務省がAI倫理に関するガイドラインを発表して以来、内閣府や経産省からも複数のガイドラインが公表されています。2024年の4月に、これらを統合するため、「AI事業者ガイドライン（第1.0版）」が策定・公表されました。

I - 1. AI倫理

<国外のAI倫理関連のガイドライン・原則>

- 「Tenets」（Partnership on AI 2016年）
- 「アシロマ AI 原則（[Asilomer AI Principles](#)）」（Future of Life 2017年）
- 「信頼できる AI のための倫理ガイドライン（Ethics guidelines for trustworthy AI）」（EU 2018年）
- 「人工知能に関する理事会勧告（Recommendation of the Council on Artificial Intelligence）」（OECD 2019年）

海外でも、様々な研究機関、NPO、学会等で多数のガイドラインが作られています。その中でも特に注目されるのが、EU の「信頼できる AI のための倫理ガイドライン (Ethics guidelines for trustworthy AI)」です。これは、2024 の 7 月にできた EU AI Act の内容に反映されており、一つのデファクトスタンダードと位置づけられています。

同様に広く普及しているのは、OECD が出している AI に関する勧告です。

I - 2. AI倫理が問題となった事例

<事例 1 : 「女性差別」 (公平性・バイアス) >

- 「女性差別」の欠陥露呈でAI採用が打ち切りになった事例
- 2018年10月、米国大手通信販売会社において、密かに準備を進めていた AI を活用した人材採用システムについて、男性を偏重するという機械学習面の欠陥が判明し、運用を取りやめる結果になったと報じられた。
- このシステムは、AI による書類選考を行うことによって、応募者をランク付けするものだったが、実際に運用してみたところ、技術関係の職種の採用過程において、性別の偏りの発生が発見されたものである。つまり、履歴書に「女性」に関係する単語、たとえば、〇〇女子大の卒業生である、

一つは「女性差別」の例です。大手の通販会社が AI を使って人材雇用を試みたところ、男性偏重の結果が出てしまい、システムの採用を断念しました。意図的に性別差別を行うためではなく、過去の採用例を学習させた結果、男性ばかり採用していたため、AI が「女性」というキーワードが含まれる履歴書を自動的に除外した事案でした。

I - 2. AI倫理が問題となった事例

<事例1：「女性差別」（公平性・バイアス）>

- （続き）あるいは「女子テニス部の副部長」といった経歴が記されていると評価が下がる傾向が確認されたのである。このようなエラーが発生した原因としては、AIの学習モデルに、同社に過去10年間にわたって提出された履歴書のパターンを学習させたところ、実際には技術職のほとんどが男性からの応募であった結果、AIは技術職には男性を採用するのが好ましいと学習したのであると推測されている。

このように不公平や差別をAIに学習させることが良いのかは難しい問題です。差別の再生産は避けるべきとする議論がある一方で、AIに現実社会の状況を変える役割を求めるべきかという声もあります。この「公平性・バイアス」の問題はまだ解決していません。

I - 2. AI倫理が問題となった事例

<事例2：再犯スコア予測（公平性・バイアス）>

- 米国の裁判所において使用されていたCOMPASという事案管理・判断予測支援のシステムでは、刑事被告人の再犯スコアを出力する機能が備わっていた。裁判所によっては、この再犯スコアを勘案して判断を下していた。
- 2016年、このシステムによる再犯スコアの算出において、高リスクと判断していたのに実際は再犯がなかったケースが黒人に高い割合で見られ、逆に低リスクと判断していたのに実際には再犯してしまったケースが白人に高い割合で見られる傾向が確認されたことが報じられ、問題となった。

「再犯スコア予測（公平性・バイアス）」も、アメリカの司法制度で問題になっています。システム導入後、算出した再犯スコアと実際の再犯率を照合すると、うまくいっていないことが判明し、このシステムは本当に適切なのかという問題になった事案です。

I - 2. AI倫理が問題となった事例

<事例2：再犯スコア予測（公平性・バイアス）>

- 実のところ、このシステムでは人種による差別が発生しないよう、黒人白人の区別は学習していなかったものの、居住地域（地域にもよるが、米国では人種の別により、居住地域がくっきりと分かれる場合が少なくない）などの他の項目をもって、事実上人種を区別できる状況であったために、このバイアスが発生したと考えられている。

また、人種差別を避けるために、黒人や白人という区別を学習させていないものの、居住地域の情報は登録していたことが仇になったケースもあります。アメリカでは都市の中心部に黒人や低収入者が住み、周辺部に高収入者や白人が住む傾向があります。これが暗黙のうちに反映され、結果的に人種を区別できるデータを学習してしまい、バイアスが発生したと指摘されています。

I - 2. AI倫理が問題となった事例

<事例3：顔認識技術による誤認逮捕（公平性・バイアス）>

- いわゆる、顔認識の「人種バイアス」問題に由来する事例である。
- 2023年8月、米ミシガン州デトロイト市で当時妊娠8カ月だった黒人女性が、AIを用いた顔認識システムによって強盗犯と誤って判断され、逮捕されてしまい、後日、女性が不当逮捕だったとして市を提訴したことが報じられた。デトロイト市警察は、監視カメラに映った強盗犯の男女の身元を割り出す過程で、AI顔認識技術を使用したところ、ある黒人女性の8年前の写真と一致するという結果が得られ、実際に逮捕に至ったものの、後日、誤認逮捕であることが判明したものである。

「顔認識の『人種バイアス』」も問題になります。監視カメラに映った強盗犯を特定するためにAIを使用したところ、誤認逮捕が発生した事案です。特に、誤認逮捕された女性が妊婦だったことが問題になりました。

I - 2. AI倫理が問題となった事例

<事例3：顔認識技術による誤認逮捕（公平性・バイアス）>

- 警察によるAI顔認識システムの欠陥が明らかになったのはこれが初めてではない。とくにデトロイト市警察はこの技術によって2019～20年に黒人男性3人を誤認逮捕しており、AIによる顔認識のアルゴリズムにはバイアスがあり、警察による人種差別を助長しているという問題の根深さが改めて浮き彫りになったとも指摘されている。このような悲劇を避けるべく、米国では、2022年9月、連邦議会においてAIの顔認識アルゴリズムの透明性を規制する初の試みとして「2022年顔認識法」案が提出されている。

結果的に、黒人が誤認逮捕されやすい状況が AI の顔認識アルゴリズムで発生し、人種差別を助長、公平性やバイアスの原則を侵害する問題になりました。これを受けて 2022 年 9 月に連邦議会で規制法案が提出されたものの、法律としてはまだ成立していません。現在、顔認識技術は世界中で広く使われ、特に捜査機関で多用されているが、ルールなしでの使用が問題視されています。特にヨーロッパでは、顔認識技術の無茶な使用を防ぐ動きが以前からあり、EU AI 規則でも規制するなど、AI 倫理から法律の規制につながった具体例が見られます。

I - 2. AI倫理が問題となった事例

<事例 4 : 顔認識による機微情報の推定 (プライバシー) >

- 2021年、スタンフォード大研究者による論文において、顔認識アルゴリズムを使って100万人の顔を分析したところ、分析対象者がリベラルなのか保守なのかという違いについて約7割の確率で判別できるという報告がなされていたことが話題になった。この原理は不明な部分が多いが、論文によれば、政治的傾向によって、カメラにまっすぐに顔を向けているかどうか、嫌悪の表情が見られるかといった部分が影響した可能性があるとの説明があった。また、同じ研究者による別の論文では、同じく顔認識アルゴリズムを用いて、性的指向も一定の確率で判別できるという報告がなされていた。

同じ顔認識でも、個人の政治的な信条や性的な趣味などを推定できるとスタンフォードの研究者が発表した事案がありました。このケースでは、このような推論に AI を使うこと自体が大きく問題になりました。

I - 2. AI倫理が問題となった事例

<事例5：フェイクニュースと社会の分断（悪用）>

- ソーシャルメディアにおける政治家アカウントのなりすましやのっとり、大量のアカウントからの偽情報の一斉発信によるフェイクニュースの問題は以前から指摘されてきたところであるが、生成 AI の登場によりディープフェイクが手軽に作成され、ソーシャルメディアで大量に流されるようになってきている。たとえば、米国のトランプ前大統領をめぐるのは、その身柄を拘束されている様子を撮影したとされる、動画生成 AI によるフェイク動画、音声生成 AI によって、同氏の声に似せた合成音声で「彼女を推薦し集会もやってあげた。だから彼女は選挙に勝てた」としゃべらせるとともに、同氏の画像を表示することで、あたかも同氏本人が告白しているかのように思い込ませるテレビ CM（共和党予備選の対立候補陣営によるもの）がすでに登場している。

次は「フェイクニュース」の事案です。2024 年のアメリカ大統領選で、トランプ前大統領やハリス候補の顔や声を無断で使用し、虚偽の動画や音声を作成してテレビやインターネットで流す行為が一般的に行われました。AI が社会の分断のために悪用されることが問題になっています。

I - 2. AI倫理が問題となった事例

<事例6：誤情報の流通（ハルシネーション（幻覚））>

- 「幻覚」とは、AI が学習したデータからは正当化できないはずの回答（=事実とは異なる内容や、文脈と無関係な内容）を堂々とする現象であり、人間が現実の知覚ではなく脳内の想像で「幻覚」を見る現象と同様に、あたかも AI が「幻覚」を見て出力しているように見えることから、このように呼ばれている。大規模言語モデルおよび基盤モデルを中核として稼働している生成 AI は機械学習をベースにしていることから、機械学習につきまとう「幻覚」（Hallucination）の問題をゼロにすることは理論上難しい。
- Google の生成 AI 「Bard」のお披露目イベントで誤った回答が示され、運営会社の株価が急落した事例

最後の事例は「ハルシネーション」の事案です。AI が事実と無関係な内容や文脈を勝手に回答してしまうことを指します。AI は確率論でそのような回答をするのが得意で、誤りが普通に起きます。例えば、Google の生成 AI のお披露目イベントで、回答が大きく間違っており、株価が下落したケースが実際にありました。

I - 2. AI倫理が問題となった事例

<事例6：誤情報の流通（ハルシネーション（幻覚））>

- 生成 AI において、「幻覚」問題に注目が集まっているのは、大規模言語モデルにより自然言語処理の能力向上の結果、回答の文章そのものが見るとまっとうに見えてしまい、内容の適否を改めて確認しない限り、回答を目にした多くの人が信じ込んでしまうこと、さらに誤解をしたまま、ソーシャルメディアによって情報が拡散されることが、現実問題として懸念されているためである。
- 人物・事業者へのいわれのない風評被害（例：国際贈収賄事件の告発者を贈賄側で有罪判決を受けた人物であるという真実とは真逆の回答が出力された事例）

特に生成 AI においては、ユーザーリテラシーが重要です。EU の AI 規則でも「AI リテラシー」が取り上げられており（第 4 条）、リテラシーの教育が不十分な中で、AI の出力情報を無断で拡散することが大問題になっています。例えば、国際的な贈収賄事件で無実の人物が有罪と誤報されたケースがあります。AI は情報の正誤を判断せずに出力するため、開発者は注意を払い、学習用データの質を高めるべきという議論が行われています。特に高度なリスクを抱える AI には、より一層注意を払うべきだということが、G7 の国際会議で議論され、共同宣言が出されました。また、EU AI 規則でも厳しい規制が課せられています。このように、AI ガバナンスに法規制をかけ、研究開発サイドにもその姿勢を共有し、問題の発生を防ぐことが求められています。

I – 3. ガイドラインに共通するポイント

<人間の尊厳・人間中心>

- 「人間中心」は、「人間の尊厳」を取る維持するが、「AI 事業者ガイドライン」では、「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（wellbeing）の追求が可能となるように行動すること」と説明する。

ガイドラインに共通するポイントの一つ目は、「人間の尊厳」や「人間中心の原則」です。人権の侵害の防止、AI を人々の手助けツールとして能力を拡張すること、多様な幸福の追求が挙げられます。

I – 3. ガイドラインに共通するポイント

<多様性の確保>

- 「信頼できる AI のための倫理ガイドライン」においては、①不公平なバイアスの回避、②アクセシビリティとユニバーサルデザインの確保、③ステークホルダーの関与が望ましいとされている。

共通するポイントの二つ目は「多様性の確保」です。EU の「信頼できる AI のための倫理ガイドライン」では、アクセシビリティやユニバーサル・デザインの確保が重要視されています。AI の多様性を確保することが強く求められています。

I-3. ガイドラインに共通するポイント

<安全性>

- 「AI 事業者ガイドライン」においては、「ステークホルダー（事業外利用者等及び第三者を含む）の生命・心身・財産、精神及び環境に危害を及ぼすことがないよう努めること」と説明されている。

共通するポイントの三つ目は「安全性」です。昨年 11 月、イギリスのブレッチリー・パークで、AI の安全性に関するサミットが行われ、各国が共同で AI の安全性の確保することが強調されました。国連の諮問機関も AI の安全性が重要性を述べ、各国で「AI Safety Institute」が設立されています。

AI はステークホルダー（利害関係者）の生命・身体・財産、精神、環境等を害する懸念があるため、このような機関の設立が進んでいます。

I-3. ガイドラインに共通するポイント

<セキュリティ>

- 「AI 事業者ガイドライン」においては、「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」と説明されている。
- 「AI 利活用ガイドライン」では、①セキュリティ対策の実施、②セキュリティ対策のためのサービス提供等、③ AI の学習モデルに対するセキュリティ脆弱性への留意を期待するとしている。

四つ目は、「セキュリティ」です。これは安全性と区別され、第三者からの攻撃への対応が議論されます。ハッキングされると偏った回答結果が生成し、人を傷つける恐れがあります。そうした事態を防ぐために、AI 事業者、特に開発やサービスの提供事業者には対策が要求されています。

I-3. ガイドラインに共通するポイント

<プライバシーの尊重>

- 「AI 利活用ガイドライン」においては、「個人情報保護法等の関連法令や各事業者のプライバシーポリシーを遵守して、社会的文脈や人々の合理的な期待を踏まえ、プライバシーが尊重され、保護されるよう、その重要性・要配慮性に応じた対応を行うこと」と定義されている。
- 「信頼できる AI のための倫理ガイドライン」においては、プライバシーの尊重のためには、適切なデータ・ガバナンスが必要であるとして、以下の遵守を求めている。

五つ目は「プライバシーの尊重」です。個人情報保護法等の文脈や各事業者のプライバシーポリシーがあります。データガバナンスの必要性が EU から強調されています。

I - 3. ガイドラインに共通するポイント

<プライバシーの尊重>

- ①収集されるデータは、違法・不公平な差別をする目的で利用されてはならない。
- ②収集されたデータは、学習に利用される前に、バイアス・誤りは訂正されなければならない。
- ③パーソナルデータを取り扱う組織においては、誰がいかなる条件でデータにアクセスできるのかを定めたプロトコルをおこななければならない。

プライバシーの尊重について、①は、違法目的や差別目的の利用をしないことが大事です。②は、収集されたデータを学習させる前にバイアス等を訂正することです。しかし、現在の大規模言語モデルではバイアス訂正は難しいです。③は、アクセス権の問題です。誰が、どのような状況でアクセスできるか明確に定めることが必要です。

I-3. ガイドラインに共通するポイント

<バイアス・公平性>

■ 「AI 事業者ガイドライン」においては、「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。」と説明されている。

■ 「AI 利活用ガイドライン」においては、AI の判断にバイアスが含まれる可能性があることに留意し、AI の判断によって不当な差別が発生することがないよう、配慮することを求めている。同ガイドラインでは、さらに学習アルゴリズムによるバイアスへの留意や人間の判断の介入が期待されるとしている。

六つ目は「バイアス・公平性」です。「AI 事業者ガイドライン」では、人種、性別、国籍、政治的信条や宗教等による偏見や差別をなくすことが重要です。また、回避できないバイアスがあることを認識しつつ、人権や多様な文化を尊重する観点から許容可能なのかを評価し、AI システム・サービスの提供や開発に活かしていくことも重要と説明されています。また、「学習アルゴリズムによるバイアスへの留意」として、差別の再生産を防ぐために人間の判断を介入させることが期待されています。

I-3. ガイドラインに共通するポイント

<透明性>

- 「AI 事業者ガイドライン」においては、「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」と説明されている。
- 「AI 利活用ガイドライン」においては、透明性に関して、AI の入出力等の検証可能性や判断結果の説明可能性に留意すべきとする。前者については、AI 提供事業者及びユーザ事業者による入出力ログを保存することが、後者については、利用者の納得感や安心感の獲得、そのためのAI の動作の証拠の提示が、求められるとする。
- プライバシーや営業秘密への配慮も同時に必要になる点が難しいとされる。

七つ目は「透明性」です。「AI 事業者ガイドライン」では、AI システム・サービスの検証可能性を確保しながら、合理的な範囲で情報を提供すべきと説明されています。「AI 利活用ガイドライン」も、AI の入出力等の検証可能性や判断結果の説明可能性に留意すべきとしています。特に入出力のログに関しては、利用者が納得できるよう証拠の提示が求められているものの、事業者にとって営業秘密に当たるため、全部開示するのが厳しいとされます。

I-3. ガイドラインに共通するポイント

<アカウンタビリティ>

- 「AI 事業者ガイドライン」においては、トレーサビリティの確保、他の 10大原則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」と説明されている。
- 「AI 利活用ガイドライン」においては、アカウンタビリティを果たす努力、AI に関する利用方針の通知・公表が求められている。
- 「信頼できる AI のための倫理ガイドライン」においては、「①監査可能性、②悪影響の最小化と報告、③適切なトレードオフ、④適切な是正」が挙げられている。

八つ目は「アカウンタビリティ」です。この言葉は論者によって、異なる意味合いで使われがちです。EU の「信頼できる AI のための倫理ガイドライン」では、監査可能性や悪影響を最小化することなどが挙げられています。

I-3. ガイドラインに共通するポイント

<アカウンタビリティ>

- アカウンタビリティについては、論者による理解が大きく異なるが、中心となる要素としては「説明可能性」「理解可能性」「追跡可能性」「答責性」「透明性」「説明責任」「補償の枠組み、救済」が挙げられている。

中心要素としては「説明可能性」、「追跡可能性」、「補償の枠組み、救済」などが挙げられています。

2. AI ガバナンスとは

ANDERSON MÖRI & TOMOTSUNE

II - 1. AIガバナンスに至るこれまでの国内の議論

AI社会実装アーキテクチャー検討会中間報告書「我が国のAIガバナンスの在り方ver. 1.0」

AI原則の例

人間中心	公正競争	教育	厚生	イノベーション	成長・持続可能性
公平性 プライバシー	公正性	安全性・セキュリティ 包摂性	透明性 頑健性	説明可能性	アカウンタビリティ 監査可能性

ハードロー
ソフトロー

↓ WhatからHow

どのようにデザインするか？

社会におけるリスク

The diagram illustrates the transition from AI principles to implementation. At the top, six principle categories are listed: Human-centered, Fair competition, Education, Welfare, Innovation, and Growth/Sustainability. Below these, specific values are grouped into ovals: Fairness and Privacy under Human-centered; Fairness under Fair competition; Safety/Security and Inclusiveness under Education; Transparency and Resilience under Welfare; Explainability under Innovation; and Accountability and Auditability under Growth/Sustainability. A downward arrow labeled 'WhatからHow' points to a grey cloud containing the question 'どのようにデザインするか？' (How to design?). To the left of this cloud is a vertical axis labeled 'ハードロー' (Hard Law) at the top and 'ソフトロー' (Soft Law) at the bottom. To the right, an arrow points to '社会におけるリスク' (Social Risks).

AI 原則は、様々なガイドラインで述べられたが、これを文章化するために作成されたのが「AI 事業者ガイドライン」です。

ANDERSON MÖRI & TOMOTSUNE

II - 1. AIガバナンスに至るこれまでの国内の議論

<AI原則→ AI原則の実装>

- AI原則の議論は、日本の『人間中心のAI社会原則』、欧州の専門家グループの『AI倫理ガイドライン』等を経て、複数国が合意したOECD AI勧告とG20 AI原則で一区切りがついたとされた。
- AI原則については概ねコンセンサスが形成されつつあるところ、テーマはAI原則から、AI原則を社会で実現するためのガバナンスの議論に移行している。

「原則」からその「実装」へと移行する過程は国によって異なります。ヨーロッパはガバナンスをハードローに落とし込み、日本はソフトロー路線を採用しました。アメリカは2023年10月に当時のバイデン大統領が大統領令を発表したほか、大手AI事業者から協力を取り付けることで合意していました。日本では、2024年の7月にAI戦略会議の下に新たにAI制度研究会が設立されました。特に、「安全性」の問題について、日本はソフトロー一辺倒では不十分という議論と見直しが行われています。

II - 1. AIガバナンスに至るこれまでの国内の議論

<基本理念>

- 人間の尊厳が尊重される社会 (Dignity)
- 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)
- 持続性ある社会 (Sustainability)

現状、ガバナンスについては、「人間の尊厳が尊重される社会 (Dignity)」「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)」「持続性ある社会 (Sustainability)」という三つの基本理念を中心に、日本国内で議論されています。

Ⅱ－１． AIガバナンスに至るこれまでの国内の議論

<AI社会原則＝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

Ⅱ－１． AIガバナンスに至るこれまでの国内の議論

<AI社会原則＝人間中心のAI社会原則（2019年3月）>

- (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

2019年3月に出された「人間中心のAI社会原則」の中に、「人間中心」、「教育・リテラシー」、「プライバシー」、「セキュリティ」、「公正競争」、「公平性、説明責任及び透明性」、「イノベーション」という7つの原則があります。

II - 2. AIガバナンスとは

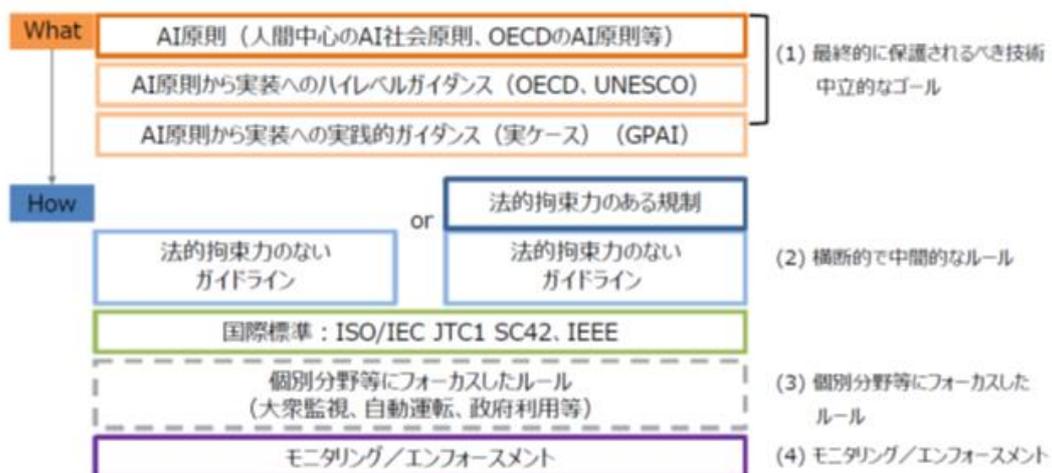
<AIガバナンスの定義>

- 「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及び運用」（経済産業省「我が国のAIガバナンスの在り方 ver1.1」）

AI ガバナンスの定義は、「正のインパクトを最大化する」ことです。AI 使用に伴うリスクを適切かつ許容可能なレベルに抑えつつ、メリットが最大化するために、技術的、組織的、社会的に検討していくことが、AI ガバナンスだと定義されています。

II - 2. AIガバナンスとは

AI社会実装アーキテクチャー検討会中間報告書「我が国のAIガバナンスの在り方 ver. 1.0」



国によって、法的拘束力のある規制を設ける場合とガイドラインにとどめる場合があります。アメリカの NIST と日本、欧州の機関が合同でその在り方を議論していま

す。また、AI を使った自動運転などの個別分野にフォーカスしたルールも発展してきます。

II-3. AIにまつわるリスク

<技術面>

- ①誤判定（例：アフリカ系米国人の画像を「ゴリラ」と誤判定する）
- ②バイアス（例：採用場面で女性応募者に低スコアをつけてしまう）
- ③誤情報（含ハルシネーション）（例：米国の裁判で弁護士が、AIによる過去の裁判例の調査結果を裁判所に提出したところ、全くのたまたまだった）
- ④偽情報（例：米国ペンタゴン付近が爆発するフェイク画像が投稿され、ダウ平均株価が急落した）

具体的に、AI にまつわるリスクは「技術面」と「社会面」に分けて分類しています。

「技術面」では、「①誤判定」、「②バイアス」、「③誤情報」、「④偽情報」が含まれます。「③誤情報」については、アメリカの裁判所での事例があります。弁護士が生成AIのチャット GPT に問い合わせた結果、ハルシネーションで誤った情報を生成し、それを裁判所に提出して懲戒を受けました。

「④偽情報」については、例えば株価に影響を与えるために、ホワイトハウスやペンタゴンが空爆された偽映像を作成し拡散することで、一瞬株価が下落し、その後に偽情報だと判明すると再び上がるため、短期間で大きな利益を得ることが可能です。

II-3. AIにまつわるリスク

<技術面>

- ⑤安全性（例：自動運転など、AI制御のシステムが物理的事故を起こす）
- ⑥セキュリティ（例：チャットボットTayに、悪意ある会話データを学習させることで、問題発言を頻発させるに至った）

「⑥セキュリティ」に関しては、チャットボットに悪意ある会話データを学習させ、問題発言を頻発させた事案が実際にありました。その結果、サービスの休止に追い込まれました。エンドユーザーがチャットボットと会話する中で問題発言を繰り返し、その内容を学習させることで、チャットボット側も問題発言がありました。

II-3. AIにまつわるリスク

<社会面>

- ①プライバシー（プロファイリング（例：米国の大手スーパー「ターゲット」が購買行動を分析して、妊娠予測スコアを算出し、クーポンを送付したところ、女子高生が含まれていた）、行動の抑制・誘導（例：映画「マイノリティ・レポート」の世界）、スコアリング（例：リクナビ、Yahoo!スコア））
- ②民主主義への悪影響（例：ソーシャルメディアの表示AIアルゴリズムによって発生する「フィルターバブル」による世論の分断）

「社会面」について、「①プライバシー」の問題 AI によるプロファイリングのリスクや行動の抑制・誘導があります。日本では「リクナビ」のように、無駄にスコアリングを行われることが問題になります。

「②民主主義への悪影響」は、分断の話題として、特に「フィルターバブル」が問題視されています。

II-3. AIにまつわるリスク

<社会面>

- ③不正利用・攻撃利用（例：ディープフェイク画像・音声・動画による、対立候補への攻撃）
- ④経済への悪影響（例：GAFAMへの資源の過度の集中、雇用の喪失）
- ⑤財産権への悪影響（例：著作物の無断学習利用、AI生成物による著作権侵害）
- ⑥電力問題（環境への悪影響）

「③不正利用・攻撃利用」と「④経済への悪影響」があります。「④経済への悪影響」は雇用の喪失や GAFAM に資源が過度に集中することが挙げられます。「⑥電力問題（環境への悪影響）」は WHO のレポートで強調され、最終的には環境への悪影響が人の生命や健康を害する可能なため、大きな社会問題です。

Ⅱ-3. AIにまつわるリスク

<リスクに影響するAIの特徴>

- ①「ブラックボックス」 対策として以下が考えられるが、十分ではない。
- ・「説明可能なAI」(Explainable AI (XAI)) の導入
- ・AIの出力結果における根拠資料の明示
- ②多数の関係当事者の存在 責任の所在が判然としない。
- ③技術革新の速度 リスクの深刻化・拡大の速度も悪化している。

リスクに影響するAIの特徴について、「①ブラックボックス」が問題です。数年前に機械学習でAIが注目を浴びた際に、AIがどのようにアルゴリズムを使っているかが開発者にも分からないまま、性能だけが向上する現象が見られました。人間はAIコントロールできないリスクが発生する懸念があります。対策として「説明可能なAIの導入」が重要視されています。AI出力結果における「根拠資料の明示」が重要です。

「②多数の関係当事者の存在」があり、責任の所在が判断しにくくなり、権利の帰属が不明確なため、紛争になりやすいです。

「③技術革新の速度」は、迅速すぎて制度が追いつかない点も問題視されています。

II-3. AIにまつわるリスク

<リスクに影響するAIの特徴>

- ④AIアルゴリズムへの信頼 AIアルゴリズムが信頼に値するかは一般にうかがい知れない。対策として、EU AI Actでは一部のAIについて検証制度をさだめるが、部分的である。
- ⑤AI倫理 トロツコ問題をはじめとする、価値のトレードオフが問題になりやすい（例：Covid-19騒動時の、街頭カメラによる、陽性者・接触者の行動追跡）
- ⑥グローバル化 AIサービスはオンラインで提供されることが多く、国レベルの規制が及び難く、及ぼそうとしても機能しない場面が多い。

「④AI アルゴリズムへの信頼」です。人間側は、AI が信頼に値するかは判断しにくいため EU AI Act の一部は開示制度や検証制度を定めました。しかし、本当に危険なものに限られたため、問題は十分に解消されないのではとされています。

「⑤AI 倫理」です。これは「トロツコ問題」のように、人によって回答が異なる場面が多く存在します。例えばコロナの時に、陽性者の行動追跡を社会的に許容されるかどうかは人によって意見が分かれました。特に AI に関する場合は、人の価値観によって結論が分かれることが多いです。

「⑥グローバル化」です。AI サービスはオンラインで提供されることが多いため、特定の国や法域で規制を強化しても、追いつきません。危険な部分に対して法律的な規制をかけても、規制の緩い国が存在する限り、完全にカバーしきれない問題が指摘されています。

II-3. AIにまつわるリスク

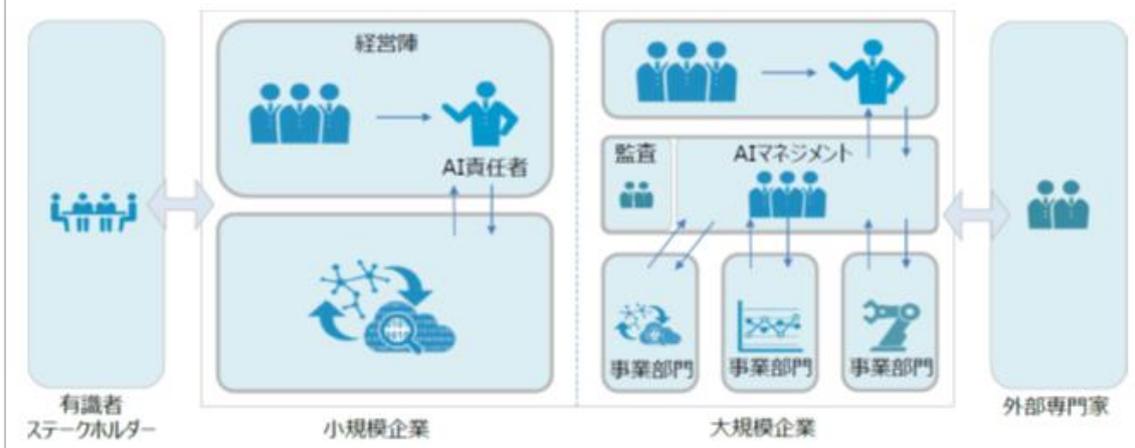
<リスクに影響するAIの特徴>

- ⑦汎用人工知能（AGI）の登場 シンギュラリティの懸念。

7つ目は「シンギュラリティ」です。

II-4. 事業者求められる関与

AI社会実装アーキテクチャー検討会中間報告書「我が国のAIガバナンスの在り方ver. 1.0」



AI事業者がどのようにAIを運用していくべきかについては、「AI事業者ガイドライン」が注意を喚起しています。ここでは事業者の規模によって異なることが指摘されています。例えば、医療分野においても、医療機関と町の小さな診療所では状況が大きく異なります。

II-4. 事業者に求められる関与

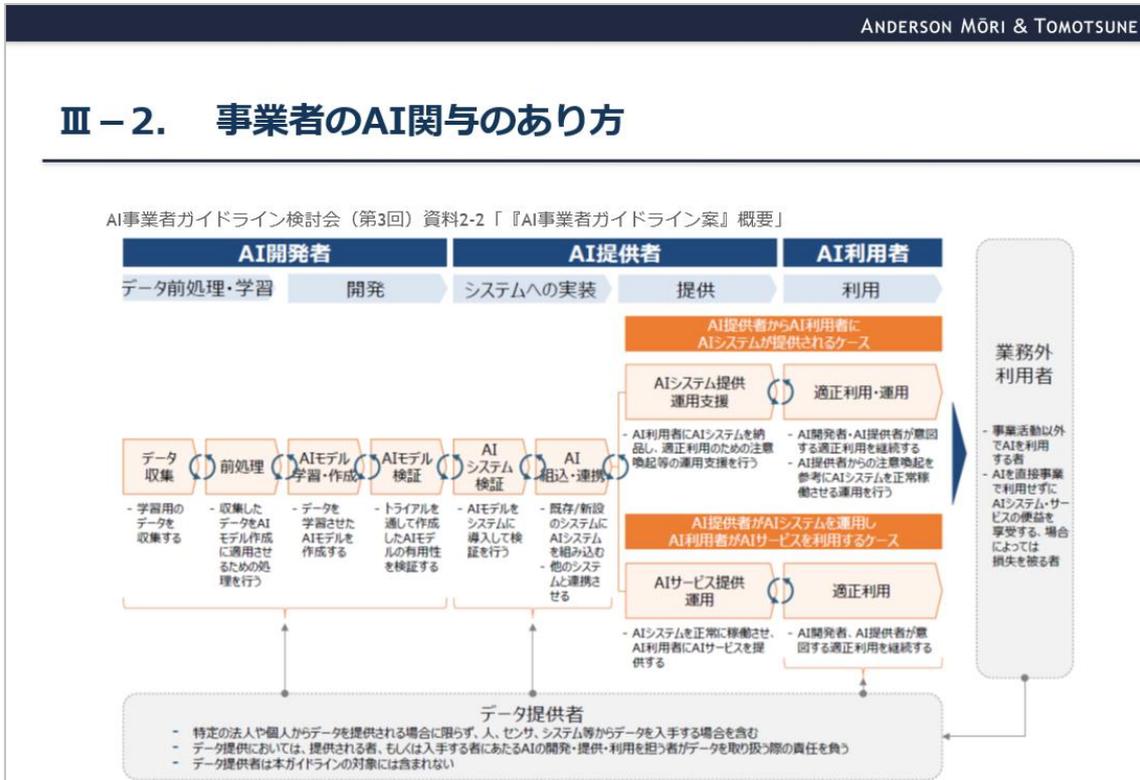
<ガバナンスの主体>

- ガバナンスの階層のレベルは、国レベル、事業者レベル、システムレベル等複数のレベルが存在しており、レベルごとに果たすべき役割は異なり、それぞれのレベルで主体的な関与が求められる。たとえば、事業者レベルで求められる役割は、国が現場に立ち会うことはできないので、事業者の主体的な関与が求められる。
- 事業者の中でも、開発者（デベロッパー）、提供者（プロバイダー）、利用者（ユーザー）によって、関係するリスクも異なることから、取り組むべきAIガバナンス対策の内容も変わってくる。

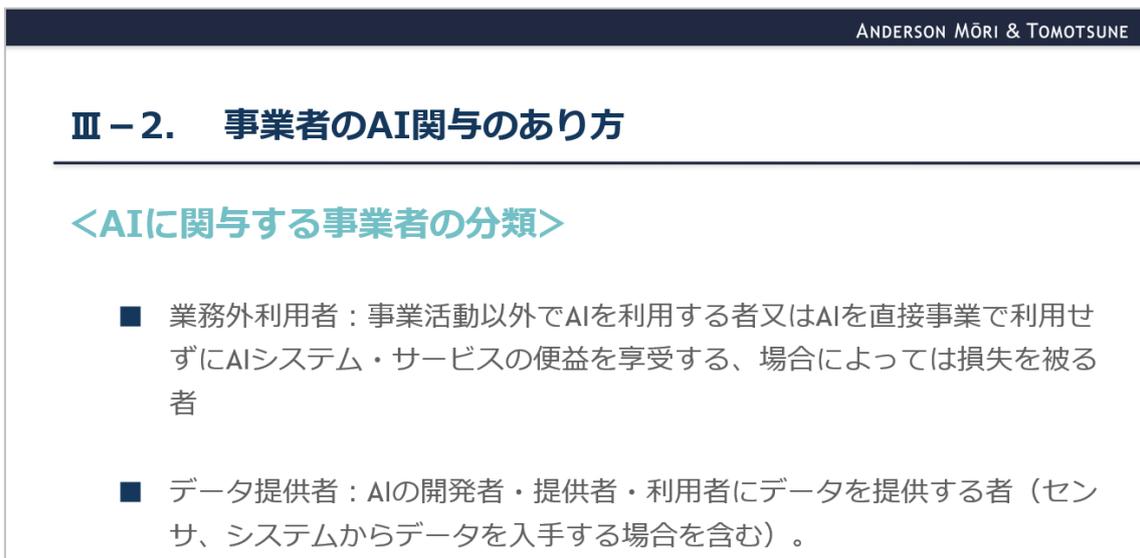
「ガバナンスの主体」については、国や事業者など様々なレベルがあり、国がすべてを実行するのは難しいため、実際には事業者が主体的に対応する必要があります。

加えて、同じ事業者内でも担当者と経営陣レベルなど、対応や役割が異なります。

3. 事業者と AI



事業者の役割は、大きく「開発者」、「提供者」、「利用者」に分類できます。



ガイドラインでは適用対象外とされているものの、ビジネスの実務を考えると、「業務外利用者」や「データ提供者」を挙げています。特に、医療の分野では、将来的に

は患者さんのデータを学習に使うことが想定されるため、「データ提供者」に該当する病院や診療所の位置づけが非常に重要です。

4. AI 事業者ガイドラインの全体像

ANDERSON MÖRI & TOMOTSUNE

IV-1. AI事業者ガイドラインの構成

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

本編 (why, what)		別添 (付属資料) (how)
主体 共通	第1部 AIとは	1. 第1部関連 [AIについて] A. AIに関する前提 B. AIによる便益/リスク
	第2部 AIにより 目指すべき社会と 各主体が取り組む 事項 A.「基本理念」 B.「原則」 C.「共通の指針」 D.「高度なAIシステムに関する 事業者に共通の指針」 E.「AIガバナンスの構築」	2. 第2部関連 [E.AIガバナンスの 構築] A. 経営層によるAIガバナンスの構築と モニタリング B. AIガバナンスの事業者取組事例
主体別	第3部 AI開発者に 関する事項 ※「高度なAIシステムを開発する組織向けの 広域プロセス国際行動規範」における 追加的な記載事項 も含む	3. 第3部関連 [AI開発者向け] A. 「第3部 AI開発者に関する事項」の解説 B. 「第2部」の「共通の指針」の解説 C. 高度なAIシステムの開発にあたって遵守 すべき事項
	第4部 AI提供者に 関する事項	4. 第4部関連 [AI提供者向け] A. 「第4部 AI提供者に関する事項」の解説 B. 「第2部」の「共通の指針」の解説
	第5部 AI利用者に 関する事項	5. 第5部関連 [AI利用者向け] A. 「第5部 AI利用者に関する事項」の解説 B. 「第2部」の「共通の指針」の解説
その他 参考資料	6. 「AI-データの利用に関する契約ガイドライン」を参照 する際の主な留意事項について 7. チェックリスト 8. 主体横断的な仮想事例 9. 海外ガイドライン等の参照先	

ガイドラインの構成本編と付属資料で大きく二つに分かれています。

5. AI 事業者ガイドラインの概要



「AI 事業者に共通して求められる指針」は、「人間中心」から「アカウンタビリティ」までの部分は各主体が取り組むべき事項として記載されています。残りの「教育・リテラシー」「公正競争確保」「イノベーション」は社会と連携して取り組む事項については、政府側が中心となることが想定されています。

V-1. AI事業者に共通して求められる指針

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

各主体が 取り組む事項	1) 人間中心	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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、各主体を含む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	<ul style="list-style-type: none"> ✓ 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

各主体が取り組むべき事項は1から10まで挙げられています。

1番の「人間中心」では、誤情報や偽情報の対策が必要とされています。また、「社会的弱者によるAIの活用」はAIが情報弱者にも利用可能であるべきとされています。

2番の「安全性」では、「学習等に用いるデータの正確性」の確保が求められています。また、データの透明性などを確保する枠組みを作ることが重要です。データの品質が低下すると、AIモデルの性能が低下し、ハルシネーションを引き起こす確率が上がるため、適切なタイミングと方法で更新することが重要です。

3番の「公平性」では、AIの出力結果が公平性を欠かないように、人間の判断を介在させることが求められています。GDPRの人間に影響を与える決定を機械任せにしないという考えと同様です。

4番の「プライバシー保護」は、事業者はプライバシーポリシー等を公開することが求められます。

5 番の「セキュリティ確保」は、一般的なセキュリティの話です。

ANDERSON MÖRI & TOMOTSUNE

V-1. AI事業者に通じて求められる指針

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

各主体が 取り組む事項 (続き)	6) 透明性	<ul style="list-style-type: none"> ✓ 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、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	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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	<ul style="list-style-type: none"> ✓ 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

6 番の「透明性」では、営業秘密やプライバシーとのバランスを考慮しつつ、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」ことが求められています。その中には、AI の利用の事実、アノテーション方法、AI システムの能力限界などの情報提供が含まれます。

7 番の「アカウンタビリティ」では、関係する情報を文書化して一定期間保管することで、エンドユーザーや末端の事業者が正しく判断できることが求められています。

8 番の「リテラシーの確保」について、医療関係における AI 情報提供など、事業者側が対応する場面も想定されます。

V-2. 高度なAIシステムに関する事業者に通の指針

<共通の指針>

- 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、アカウントビリティの向上に貢献する

広島での G7 会合の合意に基づいて作られた「高度な AI システムに関する事業者に通の指針」は AI ライフサイクル全体にわたって適切な管理を求めています。特に、透明性の確保とアカウントビリティの向上が強調されています。

V-2. 高度なAIシステムに関係する事業者に通の指針

- 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

トラブルが発生した場合には、インシデントの報告と情報共有システムの構築が求められています。不祥事や事故、トラブルの情報を業界全体や政府と共有し、対応します。

VIでは強固なセキュリティ管理を実施することが記載されています。

V-2. 高度なAIシステムに関係する事業者に通の指針

- 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

VII では電子透かし等を導入して AI 生成物であることが識別できるようにすることを求められています。これは、権利者やクリエイターの権利を守るために有効とされています。

VIII ではセキュリティ上のリスクを軽減するための研究を優先することになります。

IX は気候危機などの気候変動への対応を優先するというものです。

X は、国際的な、標準的な技術規格を作っていこうということで、実際、安全性に関して既に進められているかと思えます。

V-2. 高度なAIシステムに関する事業者に通の指針

- 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
- ※ 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。

XI ではデータインプット対策で、個人情報や著作権を保護することが求められています。

6. AI 開発者

VI-1. AI開発者特有の留意事項

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

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ることに留意し、 データの質を管理するための相応の措置を講じる - 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及び リスクを最小限に抑える 方法を検討する
	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・ガイダンスを設定し 、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ること まで検討する
	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の予測性能や品質が、活用後に大きく変動する可能性や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

「開発者特有の留意事項」として、データ前処理の段階では、プライバシー・バイ・デザインを採用し、個人情報等の適切な法措置を実施することが重要です。また、AI システムのデータの質を管理し、不公正なバイアスがないか確認することも求められています。

リスクを最小限に抑えることも重要で、問題が発生した場合には、事後検証のための記録を保存することが必要です。

VI-1. AI開発者特有の留意事項

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

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や安全性確保の仕組み、予見可能なリスクや緩和策、不具合の原因と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の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と説明 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	- 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

D-7「生じるリスク等の情報提供と説明」では、開発者から提供者に対して求められています。D-10では、研究開発等において、品質や信頼性等を確保するための措置が求められています。

7. AI 提供者

ANDERSON MÖRI & TOMOTSUNE	
Ⅶ－ 1. AI提供者特有の留意事項	
AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」	
AIシステム 実装時	<p>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</p> <p>- 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>
	<p>P-2) ii. 適正利用に資する提供</p> <p>- AI開発者が設定した範囲でAIを活用し、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>
	<p>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</p> <p>- データの公平性を担保し、参照情報、外部サービス等のバイアスを検討する</p> <p>- AIモデルの入出力や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</p> <p>- 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>
	<p>P-4) i. プライバシー保護のための仕組みや対策の導入</p> <p>- 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>
	<p>P-5) i. セキュリティ対策のための仕組みの導入</p> <p>- 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>
	<p>P-6) i. システムアーキテクチャ等の文書化</p> <p>- 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</p>

提供者に関しても「特有の留意事項」を説明されています

P-2 では、AI 開発者が想定している利用環境と実際の利用環境と齟齬がないかを、提供者が確認することが求められています。

P-3 では、バイアスの検討に加え、AI モデルの判断根拠を定期的に評価し、バイアスをチェックすることも求められています。

P-6 では、システムアーキテクチャやデータ処理プロセス等、AI システムの意思決定に影響を与える部分については、AI 提供者がきちんと文書化することが求められています。

VII – 1. AI提供者特有の留意事項

AI事業者ガイドライン検討会（第3回）資料2-2 「『AI事業者ガイドライン案』概要」

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、再発防止を検討する
	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、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、実施体制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利用者に対して、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

適正利用について。提供者は、利用者が AI システム・サービスを適正な目的で使用しているかを定期的に検証することが求められています。例えば、現在の生成 AI では、人が隠したいセンシティブな情報を工夫して引き出すこと、たとえば、芸能人が病気を隠している場合に、それを興味本位で知りたい人が AI から情報を引き出すことが可能です。このような不適切な使用を防ぐために、提供者は適切な対策を講じる必要があります。

また、P-6 では、AI を利用している事実や使用方法、更新内容等の説明の実施を求められています。

Ⅶ－２． 高度な AI システムを取り扱うAI提供者特有の留意事項

<共通の指針>

- 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システムに関する事業者に通の指針」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

「高度な AI システム」については、I から XI を適切な範囲で遵守すべきだとされています。

8. AI 利用者

Ⅷ－１． AI利用者特有の留意事項

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」

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、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、その結果が必要な関連するステークホルダーに周知 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、提供先との接点、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には、合理的な範囲で人間による判断を行い、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や要望の受付を行う
	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システムについての 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

Ⅷ-2. 高度な AI システムを取り扱うAI利用者特有の留意事項

<共通の指針>

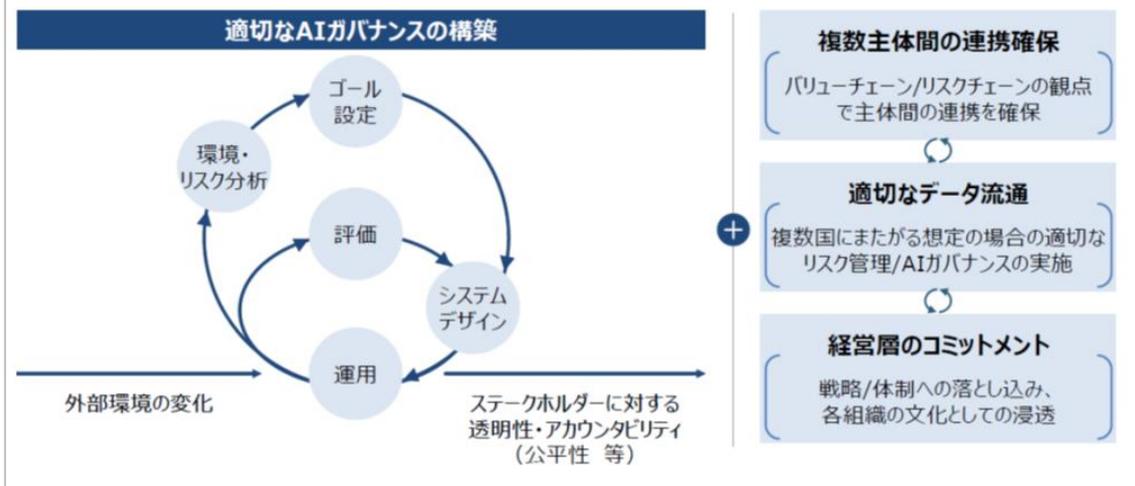
- 高度な AI システムを取り扱うAI利用者は、「第2部D. 高度なAIシステムに関係する事業者に通称の指針」について以下のように対応する。
- I) ~XI) 適切な範囲で遵守すべきである
- XII) 遵守すべきである

「高度な AI システム」に関しては、I から XI において「適切な範囲で」という表現が使われています。利用者側ができることが開発者より限られているため、このような書き方になっています。

9. AI ガバナンスの構築

IX-1. AIガバナンスの構築

AI事業者ガイドライン検討会（第3回）資料2-2「『AI事業者ガイドライン案』概要」



AI ガバナンスの構築に関しては、継続的に改善していくプロセスが求められます。

AI に関しては複数の事業者が関与することが多いため、バリューチェーンやリスクチェーンの観点から、連携を確保することが重要です。また、海外の事業者と連携する場合も多く、個人情報が含まれる場合には越境移転規制がかかる可能性があります。国によってデータの持ち出しやプログラムの輸出を制限する国もあり、規制を遵守しながらビジネスを進めるために、経営層が積極的にコミットすることが求められています。

IX-1. AIガバナンスの構築

AI事業者ガイドライン検討会（第3回）資料2-4「『AI事業者ガイドライン案』別添（付属資料）概要」

分類	行動目標 ※「3-1-1」のように更に細分化されているものもあり
1. 環境・リスク分析	1-1 便益/リスクの理解 1-2 AIの社会的な受容の理解 1-3 自社のAI習熟度の理解
2. ゴール設定	2-1 AIガバナンス・ゴールの設定
3. システムデザイン	3-1 ゴールと乖離の評価及び乖離対応の必須化 3-2 AIマネジメントの人材のリテラシー向上 3-3 各主体間・部門間の協力によるAIマネジメント強化 3-4 予防・早期対応による利用者のインシデント関連の負担軽減
4. 運用	4-1 AIマネジメントシステム運用状況の説明可能な状態の確保 4-2 個々のAIシステム運用状況の説明可能な状態の確保 4-3 AIガバナンスの実践状況の積極的な開示の検討
5. 評価	5-1 AIマネジメントシステムの機能の検証 5-2 社外ステークホルターの意見の検討
6. 環境・リスクの再分析	6-1 行動目標1-1～1-3の適時の再実施

右側が行動目標で、左側が分類です。

「環境・リスク分析」に関しては、「便益／リスクの理解」を十分に行った上で、自社のAI理解度を把握し、それを踏まえてリスク分析を行うことが求められています。

「ゴール設定」に関しては、「AIガバナンス・ゴール設定」が必要です。

「システムデザイン」では、現実とゴールの乖離、その評価をどのように行うか、またAIマネジメントをどのようにするのが課題です。更に、「運用」「評価」「環境・リスクの再分析」を行うことが重要です。

IX-1. AIガバナンスの構築

AI事業者ガイドライン検討会（第3回）資料2-4「『AI事業者ガイドライン案』別添（付属資料）概要」

別添 記載内容

解説

<p>行動目標1-1【便益/リスクの理解】： 各主体は、経営層のリーダーシップの下、AIの開発・提供・利用の目的を明確化したうえで、AIから得られる便益だけでなく見逃しにくいリスクがあることについて、各主体の事業に照らして具体的に理解し、これらを経営層に報告し、経営層で共有し、適時に理解を更新する。</p>	<p>行動目標</p> <ul style="list-style-type: none"> 事業者が取り組むことが重要となる一般的かつ客観的な目標を提示 各事業者の方針検討の際の材料となる
<p>【実践のポイント】 各主体は、経営層のリーダーシップの下、以下に取り組む。</p> <ul style="list-style-type: none"> 事業における価値の創出、社会課題の解決等のAIの開発・提供・利用の目的を明確に定義 自社の事業に結びつく形で、「便益」及び見逃せざるものを含めた「リスク」を具体的に理解 その際に、回避すべき「リスク」及び複数主体にまたがる論点に留意し、バリューチェーン/リスクチェーン全体で便益を確保、リスクを削減 迅速に経営層に報告/共有する仕組みを構築 	<p>実践のポイント</p> <ul style="list-style-type: none"> 上記の行動目標の実行のために重要となる事項や留意点を要約 各事業者が具体的な取組内容を検討する際の材料となる
<p>【実践例】 【実践例1：便益・リスクの把握】 各主体は、経営層のリーダーシップの下（担当役員又は現場に一任するのではなく、経営層自らが主導することを通して実施すること含む、以下同様）、便益だけでなくリスクについても検討し、</p>	<p>実践例</p> <ul style="list-style-type: none"> 仮想的な実践例を記載 具体的な取組イメージを持つことで、各事業者が行動につなげやすくする

更に、当事者や事業者が取り組むべき行動目標、実践のポイントと実践例も影響を受けると記載されています。

IX-2. AIガバナンスとバリューチェーン

＜バリューチェーン特有の留意事項＞

- 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

「バリューチェーン特有の留意事項」についての問題は、データの質やAIのリスクが複数の当事者に関与しているため、責任の所在が不明確になることが多い点です。仮に責任者が特定できたとしても、データや出力結果の権利帰属の問題が別途発生し、紛争の原因となる可能性があります。そのため、データの流通過程に着目し、ガバナンスの中でリスク管理を行うことが求められています。

IX-2. AIガバナンスとバリューチェーン

- 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のが想定される場合、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と、それを踏まえた相互運用性の確保（「標準」と及び「枠組み間の相互運用性」の二側面）

データの自由な流通や国境を超えた越境移転については、国によってはそれを制限する規制を設けている場合があります。このような規制の有無を含めて、ガバナンスの中で検討する必要があります。

X-1. AIによる便益の分析例

AI事業者ガイドライン検討会（第3回）資料2-4 「『AI事業者ガイドライン案』別添（付属資料）概要」

	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、新製品のプ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

こちらの表は、AIによる便益の分析例を示しています。

X-2. AIによるリスクの分析例

AI事業者ガイドライン検討会（第3回）資料2-4 「『AI事業者ガイドライン案』別添（付属資料）概要」

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	<ul style="list-style-type: none"> 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画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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 社会全体が同じモデルを、同じ温度感で使った場合、通かれる意見及び回答がLLMによって取束してしまい、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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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		<ul style="list-style-type: none"> 生成AIが事実と異なることももつとらしく回答する「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		<ul style="list-style-type: none"> 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		<ul style="list-style-type: none"> 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		<ul style="list-style-type: none"> 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		<ul style="list-style-type: none"> 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	

こちらの表は、AIによるリスクの分析例を示しています。右端に「対応する『共通の指針』」と書かれており、これはAI側が採用した1から10の指針に基づいています。

医療関係の分野では、メリットも相当大きい一方で、デメリットもある程度はどうしても避けられないかと思えます。そのため、リスクと便益の分析は医療関係でも重要になります。

XI – 1. AI開発事業者が特に留意すべき事項

<AIモデルの取扱い>

- ① ゼロから自身で創り上げる場合
- ② 外部ベンダに開発を委託する場合
- ③ 第三者から基盤モデルの提供を受ける場合
- ④ オープンソースライセンスで提供されている既存の生成AIモデルを利用する場合

ここでは、「AI 開発事業者が特に留意すべき事項」の概要を記載しています。「AI モデルの取扱い」について、権利帰属は契約や契約書によって異なるため、適切な対応が難しいです。

XI – 1. AI開発事業者が特に留意すべき事項

<その他のポイント>

- ① 生成AIモデル開発
- ② 学習用データセットの取り扱い
- ③ リスクの棚卸し作業

また、「その他のポイント」としては、「生成 AI モデル開発」「学習用データセットの取り扱い」「リスクの棚卸し作業」があります。特に学習用データセットに関しては、

大きな会社からベンダーに AI の開発を依頼し、そのベンダーが会社のデータやカスタマーデータなどを使用して学習した成果を、依頼者である会社に帰属することを認めるべきかどうかがよく問題になっています。

XI-2. AI事業者向けチェックリスト

- 人間中心の考え方を基に、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しているか？
- AIに関わる全ての者の生命・身体・財産、精神及び環境に危害を及ぼすことがないように安全性を確保しているか？
- 潜在的なバイアスをなくすよう留意し、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、回避できないバイアスが人権や多様な文化を尊重する公平性の観点から許容可能か評価しているか？
- プライバシーを尊重・保護し、関係法令を遵守しているか？

そういったことをチェックリスト化したのが、これも、「AI 事業者ガイドライン」の別添資料の中で公表されているものです。これ自体は、非常に幅が広いというか緩やかな質問なので、だいたい、イエスなどになるかなというところです。

XII-1. ヘルスケア分野のAI関連ガイドライン・報告書

<国内（順不同）>

- JaDHA「ヘルスケア事業者のための生成 AI 活用ガイド」（2024年）
- 日本プライマリ・ケア連合学会「プライマリ・ケアにおける AI 利用ガイドライン」（2023年）
- 医薬品医療機器総合機構「AI を活用したプログラム医療機器に関する報告書」（2023年）
- 科学技術振興機構「健康・医療リアルワールドデータ利活用基盤の構築と生成AIへの展開」（2023年）

ここでは、まず、ヘルスケア分野の国内のガイドラインを挙げています。

ちょうど先々週だったでしょうか、東京財団で LLM や生成 AI とヘルスケアの関連ということで、シンポジウム的なことをされていました。そちらのホームページを見ると、三つの提言のようなことが出ています。そのようなものも含めると、結構な数のヘルスケアに特化した AI 関連のガイドラインのようなものが、いま既に世の中にあるかと思います。

XII-1. ヘルスケア分野のAI関連ガイドライン・報告書

<国外>

- WHO 「AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations」 (2024年)
- OECD 「Collective action for responsible AI in health」 (2024年)
- WHO Guidance2021 「Ethics and Governance of Artificial Intelligence for Health」 (2021年)
- G7 「高度なAIシステムを開発する組織向けの国際行動規範」 (2023年)

国外に関しても、かなりの数の論文、記事等があります。その中でも重要性が高いのは、やはりこの WHO の 2024 年のドキュメントと、2021 年のドキュメント、加えて、OECD が 2024 年に出している文書です。あとは、先ほど触れた「高度な AI システムを開発する組織向けの国際行動規範」ですね。これが関係してくるところかと思っています。

XII-2. WHO LMM報告書 (2024)

WHO [AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations]

Table 1. Potential benefits and risks in various uses of LMMs in health care

Use	Potential or proposed benefits	Potential risks
Diagnosis and clinical care	Assist in managing complex cases and review of routine diagnoses Reduce the communication workload of health-care providers ("keyboard liberation") Provide novel insights and reports from various unstructured forms of health data	Inaccurate, incomplete or false responses Poor quality training data Bias (of training data and responses) Automation bias Degradation of skills (of health-care professionals) Informed consent (of patients)
Patient-guided use	Generate information to improve understanding of a medical condition (as a patient or as a caregiver) Virtual health assistant Clinical trial enrolment	Inaccurate, incomplete or false statements Manipulation Privacy Less interaction between clinicians and patients Epistemic injustice Risk of delivery of care outside the health system
Clerical and administrative tasks	Assist with paperwork and documentation required for clinical care Assist in language translation Completion of electronic health records Draft clinical notes after a patient visit	Inaccuracies and errors Inconsistent responses depending on prompts

今日はこの中で、今回のプロジェクトに一番関連性が高いと思われる、LLM でなく LMM に関する報告書をご紹介します。これは、2024 年に WHO から出されたものになります。内容自体は、ご関心があればあとで直接お読みいただくほうがいいかと思うのですが、簡単に概要だけをご紹介します。

まず、Table 1.と書いてあるところです。LMM を医療分野で活用していくときに、どのような便益、メリットと、デメリットがあるのかが整理されています。

一番上の「Diagnosis and clinical care」では、LMM を利用すると、例えば不正確であったり、不完全であったり、虚偽の反応をしてしまうといったことや、学習用データの質が低いままバイアスを余計にかけてしまうようなことになるといったこと、あるいはインフォームド・コンセントが問題になり得る、というデメリットが書いてあります。

デメリットばかりではありません。日常的な診断を AI がサポートできるようになったら、それはそれで非常にありがたいことではあるので、メリットは大きいわけですね。そのようなメリットとデメリットをちゃんと比較検討していこうということが、この報告書の Table 1.では言われています。

ANDERSON MÖRI & TOMOTSUNE		
XII – 2. WHO LMM報告書 (2024)		
WHO 「AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations」		
Medical and nursing education	Dynamic texts suited to each student's needs Simulated conversation to improve communication and to practise in diverse situations and with diverse patients Responses to questions accompanied by chain-of-thought reasoning	Contribute to automation bias Errors or false information undermine the quality of medical education New burden of learning digital skills
Scientific research and drug development	Generate insights from scientific data and research Generate text for use in scientific articles, manuscript submission or peer-review Analyse and summarize data for research Proofreading De novo drug design	Cannot hold algorithms accountable for content Algorithms encode bias towards the perspectives of high-income countries Generate information and/or references that do not exist Undermine key tenets of scientific research, such as peer review Exacerbate differential access to scientific knowledge

今回のプロジェクトに関係がありそうなのは、「科学研究と医薬品開発」、一番最後の箇所になります。これに関しては、アクセスの格差を拡大する、あるいはアルゴリズムに対してコンテンツの責任を問うことはできないので、責任の問いようがなくなるといった問題があると指摘されています。

XII – 2. WHO LMM報告書 (2024)

WHO [AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations]

Table 2. Risks to health systems associated with use of LMMs in health care

Type of risk	Description
Overestimation of the benefits of LMMs	There may be a tendency to 'technological solutionism', or over-estimation of the benefits of LMMs while ignoring or downplaying challenges in its use, including its safety, efficacy and utility.
Accessibility and affordability	Equitable access to LMMs may be lacking for several reasons, including the "digital divide" and subscription fees to access LMMs.
System-wide biases	Use of ever-larger data sets could increase biases encoded in LMMs, which could be automated throughout a health-care system.
Impacts on labour	Use of LMMs could lead to job losses in some countries and require health workers to retrain and adjust to use of LMMs. Data annotation and filtering can lead to low wages and to untreated psychological distress.
Dependence of health systems on ill-suited LMMs	Dependence on LMMs could make health systems vulnerable if LMMs are not maintained or (in low- and middle-income countries) are updated only for use in high-income countries. Furthermore, lack of preservation and protection of privacy and confidentiality could undermine trust in health-care systems by people who are not confident that their privacy will be protected.
Cybersecurity risks	Malicious attacks or hacking could undermine safety and trust in the use of LMMs in health care.

続いて、Table 2.では、医療における LMM 使用に伴う医療システムのリスクが挙げられています。ここでは、LMM を過大評価してしまうのではないかと問題や、システム全体にバイアスがかかるのではないかと書かれています。より大規模なデータセットを使うと、LMM にエンコードされるバイアスを増加させる必要が出てきてしまうので、結局医療システム全体を通して、自動的にバイアスが出てしまう可能性があるということが指摘されています。

それ以外の、労働への影響やサイバーセキュリティへのリスクなどは、必ずしも AI 特有のことではないと思うので、本日は飛ばします。

XII-2. WHO LMM報告書 (2024)

WHO [AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations]

Figure 2: Value chain of the development, provision and deployment of LMMs

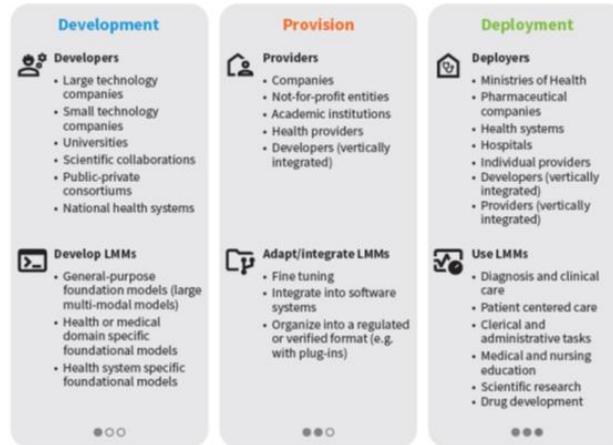


Figure 2.は、AIのバリューチェーンです。「開発者」「提供者」「利用者」はWHOのドキュメントでも、同じように「開発 (Development)」「提供 (Provision)」「展開 (Deployment)」という3種類の分類をしています。

これは、LMMとの関係での立場です。LMMの開発者は左側の青いところ、提供・統合するのは真ん中の橙色で、一番最後に使う人が出てくるということです。このようなバリューチェーンを踏まえて、WHOの報告書を読み込んでほしいという案内が、ここでされています。

XII-2. WHO LMM報告書 (2024)

WHO 「AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations」

Box 1. Brief overview of WHO consensus ethical principles for use of AI for health

- **Protect autonomy:** Humans should remain in control of health-care systems and medical decisions. Providers have the information necessary to use AI systems safely and effectively. People understand the role that AI systems play in their care. Data privacy and confidentiality are protected by valid informed consent through appropriate legal frameworks for data protection.
- **Promote human well-being, human safety and the public interest:** Designers of AI satisfy regulatory requirements for safety, accuracy and efficacy for well-defined uses or indications. Measures of quality control in practice and quality improvement in the use of AI over time should be available. AI is not used if it results in mental or physical harm that could be avoided by use of an alternative practice or approach.
- **Ensure transparency, “explainability” and intelligibility:** AI technologies should be intelligible or understandable to developers, medical professions, patients, users and regulators. Sufficient information is published or documented before the design or deployment of AI, and the information facilitates meaningful public consultation and debate on how the AI is designed and how it should or should not be used. AI is explainable according to the capacity of those to whom it is explained.

この報告書では、「健康のために AI を利用する WHO コンセンサス倫理原則」の概観が説明されています。全部は読み上げられませんが、一つ目は「自律性の保護」、二つ目は「人間の幸福、幸せぶり」という話。あとは「透明性」と「説明可能性」の話が書かれています。

XII-2. WHO LMM報告書 (2024)

WHO [AI Ethics and Governance Guidance for Large Multi-Modal Models operating in the Health Sector - Data Protection Considerations]

- **Foster responsibility and accountability** to ensure that AI is used under appropriate conditions and by appropriately trained people. Patients and clinicians evaluate development and deployment of AI. Regulatory principles are applied upstream and downstream of the algorithm by establishing points of human supervision. Appropriate mechanisms are available for questioning and for redress for individuals and groups that are adversely affected by decisions based on AI.
- **Ensure inclusiveness and equity:** AI is designed and shared to encourage the widest possible, appropriate, equitable use and access, irrespective of age, sex, gender identity, income, race, ethnicity, sexual orientation, ability or other characteristics. AI is available for use not only in high-income settings but also in low- and middle-income countries. AI does not encode biases to the disadvantage of identifiable groups. AI minimizes inevitable disparities in power. AI is monitored and evaluated to identify disproportionate effects on specific groups of people.
- **Promote AI that is responsive and sustainable:** AI technologies are consistent with the wider promotion of the sustainability of health systems, the environment and workplaces.

それから、「責任と説明責任の育成」、また「対応可能で持続可能な AI の推進」という話がここでは書かれています。

XII-2. WHO LMM報告書 (2024)

<診断と臨床治療におけるLMM使用のリスク>

- Inaccurate, incomplete, biased or false responses
- Data quality and data bias
- Automation bias
- Skills degradation
- Informed consent

少し場面を進め、「診断と臨床治療における LMM 使用のリスク」についてご紹介します。

これは、ここで挙げたとおりなのですが、先ほどのハルシネーションの問題、不正や不誠実などの原因には、そういうものが働いている可能性もあるということです。データの質と偏りの問題も、当然、指摘されています。医療情報というのは、バイアスがかかりやすい種類の情報なので、そこにきちんと気を付けた上で学習させることが必要だということが指摘されています。

それ以外には、スキルが低下してしまうリスクなどが指摘されています。

また、インフォームド・コンセントについて簡単に申し上げると、LMM の利用が増えると、対面での利用はもちろんバーチャルな利用も増えるだろうということで、AI 技術が応答を補助したり、あるいは臨床医のフィードバックを真似した生成をする可能性があることを、患者に認識してもらう必要があるとしています。ここでは、そのような意味で、インフォームドという言い方をしているようです。

XII-2. WHO LMM報告書 (2024)

<Patient-centred applicationsにおけるLMM使用のリスク>

- Inaccurate, incomplete or false statements
- Manipulation
- Privacy
- Degradation of interactions between clinicians, laypeople and patients
- Epistemic injustice
- Delivery of health care outside the health-care system

同じ話で、また少し場面が違って、これは患者を中心とするアプリケーションの場面において LMM を使用するときのリスクの一覧です。

二つ目に「Manipulation」と書いてあるところでは、ユーザーが AI を使っていくときに、変なことを教えたりしないようにということが、注意書きとして挙げられています。あとは、「Privacy」や「Degradation」などは、先ほども出てきた話かと思えますので飛ばします。

そして、分かりづらいのが「Epistemic injustice (認識論的不公正)」という言い方をしているところですが、これに関しては、いままで人間がしていた判断を AI に置き換えることにより、患者に認識論的な不公正がもたらされるということが指摘されています。

最後の「Delivery of health care」のところでは、医療システムの外で医療提供をすることについて、注意をするようにということが指摘されています。

XII – 2. WHO LMM報告書 (2024)

<Scientific and medical research and drug におけるLMM 使用のリスク>

- Lack of accountability
- High-income country bias
- Hallucination and/or misinformation
- Undermining of trust
- Accessibility of LMMs and of knowledge generated by LMMs

「Scientific and medical research and drug」に関するリスクとしては、「説明責任の欠如」、「ハルシネーション」、「信頼の低下」や「LMM によって出力された情報へのアクセシビリティが不十分であること」等が挙げられています。

XII-2. WHO LMM報告書 (2024)

<Risks to health systems and society and ethical concerns about use of LMMs>

- Overestimating the benefits of LMMs and discounting risks
- Accessibility and affordability
- System-wide biases
- Impact on labour and employment
- Dependence of health systems on unsuitable LMMs

「Overestimating the benefits of LMMs and discounting risks」については、LMM を過大評価し、リスクを軽視してしまうことが指摘されています。LMM が新しい存在であるため、十分に理解されていないことが原因です。また、LMM が手軽に利用できるかどうか、システム全体のバイアス問題も重要です。データセットの偏りにより、女性、少数民族、田舎の人、高齢者ほか、社会的に弱い立場の人のデータが除外されやすい傾向があります。これが、医療において適切なのか疑問視されています。

労働と雇用への影響ということも書いてありますが、LMM が普及すると、最終的にホワイトカラーとかブルーカラーの一部に影響があると言われているということです。

もう一つ注意すべき点は、不適切な LMM に医療が依存するのは良くないという点です。医療従事者が不足している現状では、LMM の活躍が期待されるものの、導入後にその効果を冷静に見直すことも必要だと指摘されています。

XII – 2. WHO LMM報告書 (2024)

<Laws, policies and cross-cutting requirements that could apply to use of LMMs in health care and medicine >

- Disclosure and transparency
- Data protection laws
- Assessment of general-purpose foundation models and/or applications used in health care human rights law versus risk-based frameworks
- Medical device regulation

「Laws, policies and cross-cutting requirements that could apply to use of LMMs in health care and medicine」では、法的及び政策運営の観点からリスクがあることが指摘されています。データの開示や透明性が問題となることも注意が必要です。また、「General-purpose foundation models」をヘルスケアで使用する際に、リスクベースのフレームワークで十分かどうか議論されています。EUのAI規則はリスクベースが、それを医療分野でも適応して良いかについても指摘されています。

XII-2. WHO LMM報告書 (2024)

<Risks to be addressed when deploying a health-care service or application with a general-purpose foundation model (LMM) >

- inaccurate or false responses
- bias
- privacy of data entered into and put out by an LMM
- accessibility and affordability of an LMM

「general-purpose foundation model 汎用 AI のモデル」をヘルスケアの分野で使用する際に、不適切や不正確な応答、バイアス、プライバシー侵害が懸念されます。また、十分なアクセシビリティが適切なコストで確保されているかも重要です。

XII-2. WHO LMM報告書 (2024)

<Risks to be addressed when deploying a health-care service or application with a general-purpose foundation model (LMM) >

- impacts on labour and employment
- automation bias and skills degradation
- the quality of interactions between health-care providers and patients