AI（人工知能）のガバナンス

～何を制御すべきか～

目　　次

フェロー　隅山　正敏

要　　約

Ⅰ．はじめに

企業はAIを単に利用するだけでも自主規制（ガバナンス・リスク管理）が求められる。

Ⅱ．各国政府の取組みの変遷

新技術への「懸念」を受けた各国取組みはロボット規制の検討から始まる。深層学習の登場を受けて自主規制（AI原則）を開始し、生成AIの登場を受けて自主規制の深化や法的規制の導入に転じた。

Ⅲ．米国の動向

大統領府「連邦政府によるAI利用原則」、NIST「AIリスク管理フレームワーク」、大統領府「AI版権利章典」、大統領府「安全でセキュアで信頼できるAI」について概観する。

Ⅳ．欧州連合の動向

欧州議会「ロボット憲章」、欧州倫理グループ「AI・ロボット・自律システムに関する声明」、欧州委員会「信頼できるAIのための倫理ガイドライン」について概観する。

Ⅴ．日本の動向

総務省「AI研究開発原則」、総務省「AI利活用原則」、内閣府「人間中心のAI社会原則」について概観する。

Ⅵ．国際機関の動向

G7「AI開発組織向け指導指針」、OECD「信頼できるAIのための責任あるスチュワードシップ原則」、UNESCO「AI倫理勧告」、IDCPPC「AI開発に関する指導原則」、GPA「AIシステムに関するリスク管理フレームワーク」について概観する。

Ⅶ．AI原則で使われる概念の整理

AI原則に掲げる60余りの概念は「価値／行動」、「個人／社会／技術」、「既存課題／新規課題」という評価軸で分類すると全体像を理解し易い。幅広い「懸念」をもたらすAIの「特異性」は自律・進化・汎用というキーワードで紐解くことができる。

Ⅷ．おわりに

企業（単なる利用者を含む）はAIに係る自主規制（ガバナンス・リスク管理）を求められ、その際に参照する各国政府AI原則を体系的に理解するためには上記のような分類方法が参考になる。

Ⅰ．はじめに

「ChatGPT[1]」の公開（2022 年 11 月）を転機として「人工知能」を意味する「AI（artificial intelligence）」やその一類型である「生成 AI[2]」という言葉が日常的に使われるようになった。AI は、経済成長や生産性向上の牽引役として期待が高まる一方で、誤用・悪用などによる「負の影響」を懸念する声も根強い。「実用化」の初期の段階から「期待」と「懸念」がせめぎ合う展開になっており、他の革新的な技術に見られない「特異性」が認められる。こうした「懸念」は、それを解消するための「規制」を求める声に繋がる。「規制」の要否は、過度な規制がイノベーションを阻害するという意見と、規制のもたらす安定性が社会的受入れに繋がるという意見のバランスの中で決まってくる。当事者が自らを律する「自主規制」と外部から社会的に律する「法的規制」とのいずれを選ぶのかという問題も同様である。現時点では、イノベーションを促進する観点から「自主規制」が優勢である。もっとも、「AI を搭載した自動車（自動運転車）」に道路運送車両法が適用されるといった既存の「法的規制」の上に何を加えるのかという前提での話である。

この状況に対して、殆どの企業は、「規制」が開発企業をターゲットに議論され、自社とは無関係であると認識しているであろう。しかし、AI には、これを実装した商品・サービスを利用するに過ぎない企業[3]であっても、「負の影響」の引き金を引き得るという「特異性」がある。AI が「実装後」も学習を続けるためである。したがって、製品／サービスを利用するに過ぎない企業にあっても「自主規制」（ガバナンス・リスク管理体制への組み込み）に取り組むことが求められる。

しかし、単に利用するだけの企業が「自主規制」に取り組むには困難が伴う。AI を使っているものの、理解して使っている訳ではないからである。そこで参考になるのが各国政府の提唱する「AI 原則」である。尤も、各種「AI 原則」の見出しを並べてみても、具体的な行動に落とし込むことが難しい。「見出し（概念）」の下でどのような事態が想定されているのかを掘り下げて初めて、何をすれば良いのかが浮かび上がる。そこで、本稿では各国の「AI 原則」の内容を掘り下げて、各企業が「自主規制」に取り組む際の「とっかかり」を提供することとする。

《コラム 1》人工知能のライフサイクル

人工知能は、それを搭載する製品・サービスの受渡しを終えた後も「進化」する。このため、ガバナンス・リスク管理に当たっても「ライフサイクル（設計・開発・実装・利用）の全て」をカバーする必要があるという共通認識が形成され、各国政府「AI 原則」においても「ライフサイクル」という用語が頻繁に登場する。

このことは、ガバナンス・リスク管理を行うべき主体の広がりに繋がる。総務省「報告書 2018[4]」は、人工知能の関係者として、開発者に加え「利用者（最終利用者とサービスプロバイダ）」「間接利用者（最終利用者からサービスを受ける者）」「データ提供者」「第三者（他者の AI 利用により影響を受ける者）」を挙げる。なお、間接利用者として「AI を用いたサービス（医療サービスや金融サービス）の利用者」という事例を示す。

---

[1] Open AI, *"Introducing ChatGPT"*, 2022/11/30

[2] 「generative AI」の訳で、画像の識別などに用いられる「識別 AI（discriminative AI）」と対比される。

[3] 例えば、「AI 白書 2023」（末尾参考文献）は「健康・医療・介護に対処する AI」「農業・漁業に対処する AI」「インフラ・防災・防犯に対処する AI」「交通インフラ・物流に対処する AI」「ものづくりに対処する AI」などを紹介する。

[4] <https://www.soumu.go.jp/main_content/000564147.pdf>

Ⅱ．各国政府の取組みの変遷

人工知能が「進化」すれば「懸念」も現実味を増し、各国政府の背中を押すことになる。各国政府の背中を押した「進化」と「取組内容」の変化について時系列にそって概観する。

１．先行事例としてのロボット（2010 年代前半）

人工知能に対する懸念は、その搭載製品と目された「ロボット」の時代に遡る。産業用ロボットは 1980 年代に普及期に入ったが、政府戦略に取り込まれるのは 2010 年代に入ってからである。

（１）米国のロボットに関する取組み

米国では、オバマ政権が、衰退が続いていた米国製造業の復活を目指して 2011 年 6 月に「製造業活性化戦略[5]」を発表し、その柱の一つに「国家ロボット戦略：National Robotics Initiative」を位置づけた。そこでは政府資金の投入増を決めたものの「懸念」を取り上げることはなかった。

（２）欧州連合のロボットに関する取組み

欧州連合では、欧州デジタルアジェンダ（2012 年 12 月見直し[6]）においてロボット工学を注力分野に選定する一方で、同年 3 月に「RoboLaw プロジェクト」（民間主導）を開始し、成果文書「ロボット規制ガイドライン[7]」を 2014 年 9 月にとりまとめた。この文書は、ガイドラインと銘打っているものの、規制のあるべき姿を解説するのでなく、規制の要否やその選択（どの規制を選ぶのか）を判断するポイントを整理する。規制の要否については、未熟・過剰な規制が科学の進歩を妨げるという意見と、将来を見通すことができれば安心して開発できるという意見の両論を併記する。見直すべき規制として①健康・安全・消費者・環境に関する規制、②賠償責任、③知的財産権、④プライバシーとデータ保護、⑤取引をする能力（ロボットを取引主体とすることの可否）を掲げる。規制の選択に際して考慮すべき「価値」として「平等」「連帯」「正義」「無差別」「消費者保護」などを掲げる。なお、この成果文書が後述の欧州議会「ロボット憲章」（下記Ⅳ－１）に繋がっている。

（３）日本のロボットに関する取組み

日本では、産業用ロボットで当時築いていた優位性が各国取組みにより脅かされるという危機感を背景に、成長戦略「日本再興戦略改訂 2014」（2014 年 6 月）において初めてロボットに関する取組みを取り上げた。その提言を受けて設置されたロボット革命実現会議は 2015 年 1 月に「ロボット新戦略[8]」をとりまとめ、日本経済再生本部は同年 2 月にこれを採択した。民間投資拡大に向けた政策誘導、基準・規格の開発に向けた国際的な働きかけ、ロボット利用を阻害し得る規制の緩和などを提言したが、「懸念」を取り上げることはなかった。

（４）小括

ロボットは、頭脳（人工知能）と手足（運動装置）を組み合わせた製品である。1980 年代に実用化が

---

[5] White House, *"President Obama Launches Advanced Manufacturing Partnership"*, 2011/06/24
[6] European Commission, *"Communication: The Digital Agenda for Europe – Driving European Growth Digitally"*, 2012/12/18
[7] RoboLaw, *"Guidelines on Regulating Robotics"*, 2014/09/22
[8] <https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/robot_honbun_150210.pdf>

始まり、2010 年代前半には各国の成長戦略に取り込まれるまでに至ったが、ロボットに対する「懸念」が高まることはなかった。「懸念」を取り上げた欧州ですら、「ロボット法」の制定に至らなかった。頭脳である「人工知能」がプログラムにより動かされるに過ぎず（自律性がない）、手足である「運動装置」には既存の安全規制（例えば医療機器としての規制）が適用されたため、それ以上の「懸念」が不要であったものと思われる。

２．人工知能「深層学習」の台頭（2010 年代後半）

人工知能という用語を初めて用いたのはダートマス会議の提案書（1955 年 8 月）であり、実際に開催された同会議（翌年 7-8 月）が人工知能研究の出発点とされる。その後、いくどかのブームを迎えながらも、その都度、氷河期に陥った。研究が進展しなかった訳でなく、「計算機ハードウェアの性能向上とインターネットやセンシング技術の発展によって大規模なデータが収集整備されたこと[9]」により 2010 年代にそれまでの研究成果が花開いたとされる。

（１）研究開発の動向

この「花開いた研究」が「深層学習[10]」である。Google が 2012 年 6 月に発表した、いわゆる「キャットペーパー[11]」は、人間が「猫」を教え込むことなく、機械が大量の画像を読み込むだけで「猫」を識別できるようになったと報告した。画像識別能力を競う大会（ILSVRC）において深層学習を用いたチームが他の学習方法を用いた他チームを圧倒したのも同年 10 月[12]である。人間の能力を越えるという懸念（技術的特異点＝シンギュラリティ）を意識させる出来事もこの時期に起きた。クイズ[13]（2011 年 2 月）、将棋[14]（2013 年 4 月）、囲碁[15]（2016 年 3 月）など分野が限られるものの、人工知能が人間のチャンピオンに勝利したのである。

2010 年代後半に「実用化」が始まり、これに伴って「人工知能が引き起こす問題」も生じた。Microsoft が 2016 年 3 月に公開したチャットボット「Tay」は、悪意あるユーザーの攻撃を受けて人種差別的な発言をするようになり、僅か 16 時間でサービス休止に追い込まれた[16]。サービス「開始後」も継続していた「学習」が狙われたことになる。学習データの偏りに起因する「意図しない弊害」が生じた事例もある。AI を搭載した人材採用（履歴書審査）システムを開発した Amazon は審査結果に男女差別が生じていたことを理由として 2018 年 10 月に利用を打ち切った[17]。また、クレジットカードの利用限度額の審査において男女差別の疑いがあるとしてニューヨーク州金融当局が 2019 年 11 月にゴールドマンサックスの捜査を開始した[18]。これらでは、AI システムの開発企業でなく自社用にカスタマイズした「利用企業」が非難されている。

---

9 「AI 白書 2023」43 頁。
10 深層学習の詳細については「AI 白書 2023」42-58 頁など。
11 Quoc V. Le, et al., *"Building High-level Features Using Large Scale Unsupervised Learning"*, 2012/06/26
12 総務省「情報通信白書（令和元年版）」82 頁など。
13 IBM「WATSON」が米国クイズ番組において人間のチャンピオンに勝利した。
14 第 2 回将棋電王戦において 5 つの将棋ソフトがプロ棋士に挑戦し、3 勝 1 敗 1 分けの戦績を収めた。
15 Google DeepMind「AlphaGo」が囲碁世界チャンピオンに勝利した。
16 Guardian, *"Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter"*, 2016/03/24 など。
17 ロイター「アマゾンが AI 採用打ち切り」2018/10/14 など。
18 Bloomberg, *"Viral tweet about Apple Card leads to Goldman Sachs probe"*, 2019/11/10 など。なお当局は 2021 年 3 月に違法性が認められなかったと結論付けた。

（２）各国政府の取組み

各国政府の取組みが本格化したのも、この時期になる。総務省「人工知能研究開発原則」（2016 年 4 月）、欧州科学新技術倫理グループ「人工知能、ロボット及び自律システムに関する声明」（2018 年 3 月）、米国大統領府「連邦政府における人工知能利用原則」（2020 年 12 月）などである（いずれも後述）。

３．生成 AI の登場（2020 年代）

人工知能の「第二の転換点」が冒頭に紹介した「生成 AI」である。日常で使われる言葉（自然言語）により指示（プロンプト）を出すことができるという特性を持たせて、一般人による利用を可能にし、また、後ろに控える大規模言語モデルが極めて自然な文章の作成を可能にした。一般人による利用を可能にしたことにより、利便性の向上などの「期待」と、悪用などの「懸念」が同時に高まった。

「懸念」が抽象的なものから実現味のあるものに転じたことを受けて、各国政府も取組みスタンスを変えている。米国では開発企業に対して自主規制を促す取組み（下記《コラム２》）を進め、欧州連合では法的規制（人工知能法、下記《コラム３》）に舵を切った。

Ⅲ．米国の動向

米国では、製造業（ロボット）の停滞と情報通信業（人工知能）の成長という産業構造を反映して、産業政策の重点も人工知能に置かれている。また、伝統的に「政府介入」を忌避しており、規制面でも法的規制より自主規制が選好される。

１．大統領府「連邦政府による人工知能利用原則」（2020 年 12 月）

トランプ政権は、2019 年 2 月に「国家人工知能戦略：American AI Initiative」を開始した。起点となる大統領令「人工知能分野における米国の主導権を維持する[19]」では、技術的ブレークスルーの主導など推進目標を掲げる一方で、「国民の信用・信頼を涵養する」「市民の自由・プライバシー・国民の価値を守る」という項目も立てている。ただ、「市民の自由」などの中身を掘り下げることはしていない。

こうした中、適用対象を連邦政府に限定するとはいえ「人工知能利用原則」（大統領令「連邦政府における信頼できる人工知能の利用の促進[20]」第 3 条）が 2020 年 12 月に策定された。「利用原則」は、①合法的で米国の価値を尊重、②合目的で成果志向、③正確性・信頼性・有効性、④安全性・セキュリティ・耐久性、⑤理解可能性、⑥責任と追跡可能性、⑦監視（monitor）、⑧透明性、⑨説明責任の 9 項目を掲げる。このうち「合法性」では、適用される法令・政策の全てを遵守すること（法令遵守）を求めており、プライバシー・市民権・市民の自由に関する法令を例示する。また「責任と追跡可能性」では、設計・開発・購入・利用の各局面における人間の役割と責任を定義し、理解させ、割り当てていること、それらが文書化され、追跡可能になっていることを求める。

---

19 Executive Order 13859, *"Maintaining American Leadership in AI"*, 2019/02/11
20 Executive Order 13960, *"Promoting the Use of Trustworthy AI in the Federal Government"*, 2020/12/03

２．国立標準技術研究所「人工知能リスク管理フレームワーク」（第一次案：2022 年 3 月）

国立標準技術研究所（NIST）は、測定科学・標準・技術を前進させて米国のイノベーションと産業競争力を推進する役割を持つ、商務省傘下の組織である。トランプ政権の始めた「国家人工知能戦略」の一環として NIST は 2021 年 7 月に「人工知能リスク管理フレームワーク」の開発に着手し、概念ペーパー[21]（2021 年 12 月）、第一次案[22]（2022 年 3 月）、第二次案[23]（同年 8 月）を経て 2023 年 1 月に確定版[24]（第一版）を発表している。

第一次案では、信頼できる AI として備えるべき属性を「技術的なもの」「社会・技術の相関で生まれるもの」「守るべき価値（指導原則）」の 3 つに区分する。技術的属性として①正確性、②信頼性、③頑健性、④耐久性・セキュリティを、社会・技術的属性として⑤説明可能性（何が出力を決めたか）、⑥解釈可能性（何が出力に影響したか）、⑦プライバシー、⑧安全性、⑨バイアス管理を、指導原則として⑩公正性、⑪説明責任、⑫透明性を掲げる。

確定版（第一版）では、3 つの区分を廃し、①妥当性（必要要件の充足）と信頼性、②安全性、③セキュリティと耐久性、④説明可能性・解釈可能性、⑤プライバシー、⑥公正性（偏見の管理）、⑦説明可能性・透明性という 7 項目を同列に並べる。他の AI 原則がリスクの中身や対応要領を説明するのに対し、本文書では各要件の定義（例えば、信頼性とは要求どおりに失敗なく機能することであると定義する）を掲げる点に特徴がある。

３．大統領府「人工知能版権利章典」（2022 年 10 月）

「権利章典：Bill of Rights」は、元々イギリス名誉革命時に国王と議会とで権限を分配する目的で制定された法律であるが、米国では合衆国憲法のうち人権保障規定を指す用語として用いられる。その AI 版は、AI を含む自動化システムが人々の監視やランク付けに用いられるなど、人々の権利を脅かすようになっているという問題意識に基づく。科学技術政策局は大統領令[25]を受けて 2021 年 10 月に検討に着手し、2022 年 10 月に「人工知能版権利章典の青写真[26]」として発表した。そこでは①安全性・有効性、②アルゴリズムによる差別からの保護、③データプライバシーの確保、④通知と説明、⑤人間による対応を選ぶ機会の付与という 5 項目を掲げる。このうち「通知と説明」では、自動化システムが個人に係る事項に関与する場合、その人は、システムを利用している旨の「通知」と、決定に至る手順などの情報に関する「説明」を受けるべきであるとする。

４．大統領令「安全でセキュアで信頼できる人工知能」（2023 年 10 月）

ChatGPT の急速な普及を受けて、バイデン政権は 2023 年 5 月に「責任ある人工知能イノベーションを促進する行動[27]」を発表するなど、取組みを進めている（《コラム 2》参照）。そうした中、2023 年 10

---

[21] NIST, *"AI Risk Management Framework Concept Paper"*, 2021/12/13
[22] NIST, *"AI Risk Management Framework: Initial Draft"*, 2022/03/17
[23] NIST, *"AI Risk Management Framework: Second Draft"*, 2022/08/18
[24] NIST, *"AI Risk Management Framework (AI RMF 1.0)"*, 2023/01/26
[25] White House, *"Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government"*, 2021/01/20
[26] White House, *"Blueprint for an AI Bill of Rights"*, 2022/10/04
[27] White House, *"Fact Sheet: Biden-Harris Administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety"*, 2023/05/04

月に大統領令「安全でセキュアで信頼できる人工知能[28]」を発出した。研究の推進、専門人材の受入れ、医療・福祉・教育分野での活用の推進などの推進政策を打ち出す一方で、安全・セキュリティの確保、労働市場に対する影響への対処、衡平（equity）と市民権の推進、プライバシーの保護にも措置を講じる。安全面では、安全性テスト結果を政府と共有する、安全性等の標準を開発する、有害なバイオ物質の製造を阻止する、人工知能により可能になる詐欺等に対処するという公約を掲げる。また、労働関係では労働市場に対する影響、失職者に対する支援に係る報告書の作成を、「衡平」では差別防止のためのガイダンスの策定、刑事司法制度における好ましい利用事例の収集などを、プライバシーでは連邦政府向けガイダンスやプライバシー保護技術を評価するガイダンスの策定をそれぞれ指示する。

５．小括

これらの「AI 原則」は、様々な「概念」を使って「AI のもたらす課題（デメリット）」への備えを呼びかけている。デメリットを誰が被るのか（対象者）という観点から「個人にとっての課題」「社会にとっての課題」「技術上の課題」に分けることができる（下記Ⅶ－２）。この 3 つの区分に属する 18 の概念に基づいて、これら「AI 原則」を比較したのが《図表 1》である。当初は安全性・セキュリティなどの技術的課題を懸念し、次第に個人的課題にシフトする様子が窺える。また、社会的課題への関心は薄い状況が続いている。

《図表 1》米国のＡＩ原則

区分	項目	政府原則 2020/12	ﾘｽｸ管理 2022/03	権利章典 2022/10	大統領令 2023/10
個人	人間の尊厳				
	人権の尊重	△			○
	プライバシー	△	○	○	○
	公正性・公平性		○		
	平等（無差別）		○	○	○
	自律性・自己決定権			△	
社会	民主主義				
	法の支配				
	包摂性				
	多様性				
	広義の責任	○			
	賠償責任				
技術	安全性	○	○	○	○
	セキュリティ	○	○		○
	無害性・危害防止				
	透明性	○	○		
	説明責任	○	○	○	
	説明可能性		○	○	

（注 1）比較する原則は左から「連邦政府における人工知能利用原則」「人工知能リスク管理フレームワーク」「人工知能版権利章典」「大統領令：安全でセキュアで信頼できる人工知能」である

（注 2）「○」は見出しに使われる概念を、「△」は説明文の中に登場する概念を表す

（出典）当社作成

《コラム 2》米国政府による自主規制の推進

バイデン政権は、生成 AI の急速な普及を受けて 2023 年 5 月に「責任ある AI イノベーションを促進する行動[29]」を発表し、研究開発投資の追加、生成 AI に関する評価の実施、政府利用に関する指針の策定を表明した。このうち「生成 AI 評価」の成果として 2023 年 7 月に主要 7 社から自主的な取組

[28] Executive Order 14001, *"Safe, Secure, and Trustworthy AI"*, 2023/10/30

[29] White House, *"Fact Sheet: Biden-Harris Administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety"*, 2023/05/04

みを取り付け[30]、同年 9 月には対象企業を 8 社追加した[31]。自主的な取組みは、安全性、セキュリティ、信頼性という 3 つの原則に基づく 8 項目からなる。安全性に関してはバイアスなどに関するテストをシステム公開前に実施すること、リスク管理に関する情報を産業界・政府・学会などと共有することを約束し、セキュリティに関してはサイバーセキュリティ対策や AI 開発に関わる知的財産の保護などを約束し、信頼性に関しては AI 生成コンテンツであることを利用者が認識できるように電子透かしなどの仕組みを開発することなどを約束する。2024 年 2 月には生成 AI の安全な開発・実装を支援する「AI 安全研究所コンソーシアム（AISIC）」の参加企業 200 社超を発表した[32]。

Ⅳ．欧州連合の動向

欧州連合は、強い製造業（ロボット）と劣位にある情報通信業（人工知能）という産業構造を反映して、ロボット（ハード）の取組みが先行し、人工知能（ソフト）の取組みに影響を与えている。また、研究開発の推進（アクセル）と規制の検討（ブレーキ）を並行的に進め、法的規制を自主規制より重視する傾向が見られる。

１．欧州議会「ロボット憲章」（2017 年 2 月）

欧州議会の法務委員会は、「RoboLaw プロジェクト」（上記Ⅱ－１）の報告を受けて、2015 年 1 月に作業部会を設置した。作業部会がとりまとめた報告書（決議案）は、欧州議会において 2017 年 2 月に採択された。採択された「ロボットに関する民事ルールに関する決議[33]」は、欧州委員会に対して研究投資の拡大、倫理問題に関する検討、専門機関の新設などを要請するものであるが、その中に「ロボット憲章」の提案が含まれている。提案する憲章は「技術者のための倫理行動規範」「研究倫理委員会のための規範」「設計者向けライセンス」「利用者向けライセンス」により構成される。そのうち行動規範は、倫理的行動の要求と基本原則の遵守からなり、倫理的行動として「有益性」「無害性」「自律性」「正義（便益の公正配分）」に基づく行動を求め、また、基本原則として①基本権、②予防性（予防原則[34]）、③包摂性（情報提供と意思決定への全員参加）、④説明責任、⑤安全性、⑥可逆性（いざというときに元に戻す仕組み）、⑦プライバシー、⑧利益最大化・被害最小化を掲げる。なお、利用者ライセンスは、他人のプライバシーの尊重、兵器への転用の禁止など利用が許諾される条件を定める。

２．欧州科学新技術倫理Ｇ「人工知能・ロボット・自律システムに関する声明」（2018 年 3 月）

「欧州科学新技術倫理グループ：EGE」は、研究・技術開発政策に倫理的配慮を埋め込む目的で欧州委員会が 1991 年 11 月に設置した独立諮問機関であり、意見書「ヒト幹細胞の研究と使用の倫理的側

[30] White House, *"Fact Sheet: Biden-Harris Administration secures voluntary commitments from leading AI companies to manage the risks posed by AI"*, 2023/07/21、など。
[31] White House, *"Fact Sheet: Biden-Harris Administration secures voluntary commitments from eight additional ai companies to manage the risks posed by AI"*, 2023/09/12
[32] US Department of Commerce, *"Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety"*, 2024/02/08
[33] European Parliament, *"Resolution on Civil Rules on Robotics"*, 2017/02/16
[34] 通常の規制は「結果」との因果関係が特定された「行為」を規制するが、環境規制などでは因果関係の特定作業を待つことなく「結果をもたらす可能性のある行為」を規制して「結果発生」を予防するという考え方（予防原則）がとられる。

面」（2000 年 11 月）などを発表してきた。同グループは 2018 年 3 月に「人工知能、ロボット、自律システムに関する声明[35]」を発表した。声明は、人工知能等の設計・製造・利用・管理における倫理的・法的枠組みを整備すべきであるという立場から、そのたたき台となる 9 つの倫理原則を提示した。具体的には①人間の尊厳、②自律、③責任、④正義・衡平・連帯、⑤民主主義、⑥法の支配と説明責任、⑦セキュリティ・安全・心身の一体性、⑧データ保護とプライバシー、⑨持続可能性である。

このうち「人間の尊厳」では交信相手が機械なのか人間なのかを知ることが重要だとする。また、「責任」とは人工知能等がもたらす影響に責任を持つこと、「正義」とは便益を公正に分配すること、「連帯」とは社会保障などの相互扶助システムを壊さないこと、「説明責任」とは法の支配を実現するメカニズムについて説明することであるとする。分かり難いのが「心身の一体性（bodily and mental integrity）」であり、そこでは、人工知能等が心身の一体性に関する権利を損なわないよう事前テストを実施すべきであるとする。

３．欧州委員会「信頼できる人工知能のための倫理ガイドライン」（2019 年 4 月）

欧州委員会は、デジタルサービス市場が米国企業に席巻されているという危機感から開始した「デジタル単一市場戦略[36]」（2015 年 5 月）の中間レビュー[37]（2017 年 5 月）において、人工知能に初めて言及し、これが政策文書「欧州のための人工知能[38]」に繋がった。そこで提示された構想を具体化するために 2018 年 6 月に設置された作業部会は、2018 年 12 月に「信頼できる人工知能のための倫理ガイドライン[39]」を発表し、意見公募手続を経て 2019 年 4 月に確定させた。ガイドラインは「3 つの要素」「5 つの基本権」「4 つの倫理原則」「7 つの要件」という複層構造をとる。

（１）3 つの要素

「信頼できる人工知能」であるためには①合法的であること、②倫理的であること、③頑健であることという 3 要素を満たす必要があるとする。また、当然の前提となる「法令遵守＝①」を解説の対象から外し、②と③を満たすために遵守すべき倫理原則（基本権から抽出している）と、これらを実現するために実行すべき要件に分けて解説する。

（２）5 つの基本権

倫理原則の基礎になる基本権として「人間の尊厳」「個人の自由」「民主主義・正義・法の支配」「平等・無差別・連帯」「市民権」を掲げる。このうち「正義（justice）」と「連帯（solidarity）」に明確な説明を付していないが、正義は「民主的プロセスや（機械でなく）人間による熟慮を没却させない」という意味で、連帯は「弱者を見放さない」という文脈で用いられている。また「市民権」は政府が AI を公共サービスに利用する際の「負の影響」から市民を守るという文脈で用いられている。

---

[35] European Group on Ethics in Science and New Technologies, *"Statement on Artificial Intelligence, Robotics and Autonomous Systems"*, 2018/03/09

[36] European Commission, *"Communication: A Digital Single Market Strategy for Europe"*, 2015/05/06

[37] European Commission, *"Communication on the Mid-Term Review on the Implementation of the Digital Single Market Strategy"*, 2017/05/10。「第 4 章：デジタルトランスフォーメーションの管理」の中で「AI 能力を築く」という項を立てる。

[38] European Commission, *"Communication: AI for Europe"*, 2018/04/25

[39] High-Level Expert Group on AI, *"Ethics Guidelines for Trustworthy AI"*, 2019/04/08

（3）4 つの倫理原則

倫理原則として「人間の自律」「危害の防止」「公正性」「明瞭性（explicability）」を掲げる。このうち「自律」では、自らに関する決定権、民主的プロセスへの参加権、選択の機会を人間から奪ってはならず、AI システムは人間を不当に従属・強要・欺罔・操作してはならないと説明する。「危害防止」では安全性・セキュリティ・技術的頑健性を持つべきで、悪意ある利用を排除し、弱い人々に対する配慮と参加（いわゆる包摂）を確保すべきとする。「公正性」では「便益と費用の平等・公正な配分」や「バイアス・差別からの自由」を確保すること、「AI システムの決定を争い、救済を求める」手続を整備することを求める。「明瞭性」の定義として、プロセスが透明で、AI システムの目的・性能が広く伝わっており、その決定が説明可能であることを掲げる。また、説明が不可能である場合には他の方法（追跡可能性、監査可能性、透明なコミュニケーション）を探るべきだとする。

（4）7 つの要件

倫理性と頑健性を実現する要件として(a)人間の代理と監督、(b)技術的頑健性・安全性、(c)プライバシーとデータガバナンス、(d)透明性、(e)多様性・無差別性・公正性、(f)環境・社会の福利、(g)説明責任を掲げる（《図表 2》）。「代理（agency）」とは、人間を代理して動く人工知能のことで、人間が頭で考えると人工知能が義足を動かすといった使用事例を念頭に「human agency を支援すべき」とする。

《図表 2》倫理ガイドライン「7 つの要件」

要件	主な記載事項
代理と監督	・基本権を脅かすリスクがあるときは影響度評価を実施すべきである ・AI システムを理解し、安心して利用するための知識とツールを提供すべきである ・人間の自律性を損なわないために人間による監督を導入すべきである
頑健・安全	・頑健性とは意図どおりに振る舞うこと、予期しない危害を最小化することである ・敵対者が狙う攻撃目標はデータ（データ汚染）、モデル（モデル流出）、基盤である ・問題発生時のフォールバックプランを用意すべきである
プライバシー	・プライバシー保護は利用者が最初に提供する情報だけでなく、利用時に生成される情報をもカバーすべきである ・学習データに含まれるバイアスなどには学習前に対処すべきである
透明	・AI システムが下す決定についてデータセットとプロセスを文書化し追跡可能にすべきであり、それが監査可能性・説明可能性に繋がる ・利用者は相手が AI システムであることを知る権利を持つ
多様性など	・ライフサイクルの各段階で包摂性（抜け落ちる人がいない）・多様性（背景・文化・規範の異なる人が参画する）を実現すべきである ・B2C 分野で利用される AI システムは弱者でもアクセスできるようにすべきである
福利 Well-being	・公正の原則・無害の原則は他の生物や環境にも適用すべきである ・持続可能性や生態に関する責任を負い、地球規模の課題の解決に資するべきである ・利用者があらゆる生活領域で AI システムに接するようになると様々な考えが変わる
説明責任	・開発・利用の前後を通じて AI システムとその出力に対する責任 responsibility・説明責任 accountability を確保するメカニズムを導入すべきである

（出典）当社作成

4．小括

欧州連合における 3 つの「AI 原則」を比較してみると（《図表 3》）、ロボット憲章が安全性・無害性

などの「技術的課題」に重点を置いたのに対し、AI 原則は「個人的課題」への関心を高めている。また、「社会的課題」にも配慮しており、やや総花的な印象も受ける。

《図表 3》欧州連合のＡＩ原則

区分	項目	ﾛﾎﾞ憲章 2017/02	倫理声明 2018/03	倫理GL 2019/04
個人	人間の尊厳	△	○	○
	人権の尊重	○		○
	プライバシー	○	○	○
	公正性・公平性		△	○
	平等（無差別）		○	○
	自律性・自己決定権	○	○	○
社会	民主主義		○	○
	法の支配		○	○
	包摂性	○	△	△
	多様性			○
	広義の責任		○	△
	賠償責任		△	
技術	安全性	○	○	○
	セキュリティ		○	△
	無害性・危害防止	○		○
	透明性	△	△	○
	説明責任	○	○	○
	説明可能性			

（注 1）比較する原則は左から「ロボット憲章」「人工知能、ロボット及び自律システムに関する声明」「信頼できる人工知能のための倫理ガイドライン」である

（注 2）「○」は見出しに使われる概念を、「△」は説明文の中に登場する概念を表す

（出典）当社作成

《コラム 3》人工知能法

人工知能法案は、欧州委員会の提案（2021 年 4 月）、欧州理事会の修正（2022 年 11 月）、欧州議会の修正（2023 年 6 月）を経て、2023 年 12 月に政治合意に至っている。今後、理事会・議会の正式採択を経て公布・施行となる。

法案は、AI システムをその抱えるリスク（容認できないリスク、高リスク、限定リスク、最小リスク）に応じて区分し、容認できないリスクを抱える AI システムを禁止するとともに、その他の AI システムにリスクに比例した義務を課す。禁止される AI システムとしてサブリミナルな技法を用いるもの、集団の脆弱性を利用するものなどを掲げる。高リスクのシステムにはリスク管理・データガバナンスに関する枠組みの構築や事前の適合性評価の実施などを義務付ける。限定リスクのシステムには利用者に対する情報提供（AI システムが対応している旨の通知など）を義務付け、最小リスクのシステムには特段の義務付けを行わず、自主的な対応のみを求める。また、生成 AI などの汎用 AI（GPAI: general purpose AI）について義務を加重する。

Ⅴ．日本の動向

日本は、産業構造の面では欧州連合に近いものの、自主規制の選好という面では米国に近い。また、倫理面の検討は、ロボットに関しては行われず、人工知能に関する検討では世界に先駆けている。

１．総務省「人工知能研究開発原則」（2016 年 4 月）

総務省は、2015 年以降、情報通信行政の立場から人工知能に取り組んでいる。「インテリジェント化が加速する ICT の未来像に関する研究会報告書[40]」（2015 年 6 月）では、情報通信技術（ICT）が認知や判断といった能力を獲得するようになると、その研究・開発原則を明らかにし、発生し得る負の側面を小さくする仕組みを構築する必要が生じるとする。研究・開発原則については、「人工的な知性の行動を最後は人間が制御可能とすること、サイバー攻撃やセンサ攪乱攻撃に対して十分な耐性を確保すること、プライバシー保護を確実にすること、リスク分析と管理を行うこと等が考えられる」とする。

上記研究会の後継組織「AI ネットワーク化検討会議」は、その中間報告書[41]（2016 年 4 月）において「研究・開発原則に盛り込むべき事項」を整理する[42]。具体的には①透明性の原則（説明可能性・検証可能性）、②利用者支援の原則（選択の機会の提供を含む）、③制御可能性の原則、④セキュリティ確保の原則（頑健性・信頼性）、⑤安全保護の原則（生命・身体の安全の確保）、⑥プライバシー保護の原則、⑦倫理の原則（人間の尊厳と個人の自律）、⑧アカウンタビリティの原則（情報提供とコミュニケーション）を掲げる。この内容は、同月開催の G7 情報通信大臣会合において国際的議論のたたき台として披露された。

２．総務省「人工知能利活用原則」（2018 年 7 月）

検討会議の後継組織「AI ネットワーク社会推進会議」は、報告書 2018 [43]（2018 年 7 月）において「利活用原則」を案として提示する。AI システムが利用の局面に入ってからも学習等により出力やプログラムを変化させることから、利用者・データ提供者を対象とする原則が必要であるとの認識に基づく。具体的には①適正利用の原則（役割分担の明確化）、②適正学習の原則（学習データの質の確保）、③連携の原則（システム相互間の連携）、④安全の原則、⑤セキュリティの原則、⑥プライバシーの原則、⑦尊厳・自律の原則、⑧公平性の原則、⑨透明性の原則（入出力の検証可能性と判断結果の説明可能性）、⑩アカウンタビリティの原則を掲げる。「研究・開発原則」と比べると、(a)利用者支援の原則を適正利用・適正学習の原則に置き換える、(b)制御可能性の原則を外し、連携の原則を加える、(c)公平性の原則を加え、倫理の原則の表題を尊厳・自律の原則に変えるといった点を指摘することができる。

３．内閣府「人間中心の人工知能社会原則」（2019 年 3 月）

内閣府は、2016 年以降、科学技術行政の立場から人工知能に取り組んでいる。第 5 期科学技術基本計画（2016 年 1 月）において目標に掲げる「超スマート社会」の基盤技術の一つに AI 技術を掲げるとともに、並行して「倫理的・法制度的・社会的課題」の検討が必要であるとする。これを受けて設置され

---

40 <https://www.soumu.go.jp/main_content/000363712.pdf>
41 <https://www.soumu.go.jp/main_content/000414122.pdf>
42 AI ネットワーク社会推進会議「報告書 2017」において「研究・開発原則」を一部改定している。
43 <https://www.soumu.go.jp/main_content/000564147.pdf>

た「人工知能と人間社会に関する懇談会」は報告書[44]（2017 年 1 月）において、AI 技術が人間の選択や判断を支援する場合、人間の心や行動を操作・誘導したり、評価・順位付けをしたり、感情・愛情・心情に働きかけたりする場合に倫理的検討が必要になるなど、倫理的・法的論点、経済的論点、教育的論点、社会的・研究開発的論点について解説している。また、AI を受け入れる社会のあり方を検討するために「人間中心の AI 社会原則検討会議」を設置し、その報告書[45]（2019 年 3 月）において「AI 社会原則」を提唱する。そこでは①人間中心の原則（基本的人権の尊重、依存・悪用の排除）、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、⑥公平性・説明責任・透明性の原則（公平性・透明性のある意思決定、その結果に対する説明責任、技術に対する信頼性）、⑦イノベーションの原則という 7 項目を掲げる。

《図表 4》日本の AI 原則

区分	項目	開発原則 2017/07	利用原則 2018/07	社会原則 2019/03
個人	人間の尊厳	△	○	△
	人権の尊重			△
	プライバシー	○	○	○
	公正性・公平性		○	○
	平等（無差別）	△		△
	自律性・自己決定権	△	○	△
社会	民主主義			
	法の支配			
	包摂性			
	多様性			△
	広義の責任			△
	賠償責任			
技術	安全性	○	○	
	セキュリティ	○	○	○
	無害性・危害防止	△		
	透明性	○	○	○
	説明責任	○	○	○
	説明可能性	△	△	

（注 1）比較する原則は左から「人工知能研究開発原則」「人工知能利活用原則」「人間中心の人工知能社会原則」である
（注 2）「○」は見出しに使われる概念を、「△」は説明文の中に登場する概念を表す
（出典）当社作成

4．小括

以上の「AI 原則」を比較してみると（《図表 4》）、開発原則が「技術的課題」に重点を置くのに対し、利活用原則が「個人的課題」までカバーしており、遵守を求める対象者の違いが表れている。また、米国同様に「社会的課題」に対する関心は薄い印象である。

Ⅵ．国際機関の動向

人工知能を搭載したサービスは、他のデジタルサービスと同様に、国をまたがって提供されるため、比較的早い段階から国際的な議論が進められている。関係する国際機関の取組状況を概観する。

1．G7「人工知能開発組織向け指導指針」（2023 年 10 月）

（1）以前の取組み

G7 首脳会議における合意文書を概観すると、2017 年 5 月「イノベーション・技能・労働に関する G7

[44] <https://www8.cao.go.jp/cstp/tyousakai/ai/summary/aisociety_jp.pdf>
[45] <https://www8.cao.go.jp/cstp/aigensoku.pdf>

人間中心の行動計画[46]」において「包括的な成長を生み出すドライバーとしてAIの開発と利用を促進する」という言及がなされ、2018年6月「AIの未来のためのシャルルボア共通ビジョン[47]」に繋がった。そこでは「人間中心のAIの推進」「AI投資の促進」と並び、「多様な人々の参画の確保」「AIの信頼を促進する努力の支援」に言及する。

（2）指導指針の策定

2023年5月の広島サミットでは生成AIのガバナンスの枠組みを検討する「広島AIプロセス」で合意し[48]、同年10月に「開発組織向けの指導指針と行動規範[49]」を追認した。指導指針は関係者が遵守すべき項目を列挙したものであり、行動規範は指針を遵守した行動をとる上での手引きである。指導指針では、①リスクの特定・評価・軽減、②誤用例の把握、③性能・限界・使用領域の開示、④インシデント情報の共有と報告、⑤ガバナンス方針・リスク管理方針の開示、⑥セキュリティ管理、⑦AI生成コンテンツを識別する仕組み作り、⑧社会面・安全面・セキュリティ面のリスクの軽減、⑨世界的課題を解決するAIシステムの開発、⑩国際的技術規格の開発、⑪個人データ保護と知的財産権に配慮したデータ入力の確保という実践的な項目を掲げる。行動規範は、指導指針に掲げる11項目を行動に落とし込んでいる。その際にリスクの中身を分析しており、(a)化学・生物学・放射線・原子力リスク、(b)サイバー攻撃、(c)健康・安全上のリスク、(d)複製リスク、(e)社会的リスク（バイアス・差別・プライバシー・データ保護など）。(f)民主的価値・人権に対する脅威、(g)連鎖反応をもたらすリスクを掲げる。

2．OECD「信頼できる人工知能のための責任あるスチュワードシップ原則」（2019年5月）

OECDは、G7香川・高松情報通信大臣会合（2016年4月）を契機としてAIに関する取組みを開始した。AIに関する理事会勧告を策定するために2018年5月に専門家会合を設置し、その成果を踏まえて、閣僚理事会は2019年5月に勧告[50]を採択した。勧告は「AI原則」と「各国への提言」からなり、前者の正式名称は「信頼できるAIのための責任あるスチュワードシップに関する原則」である。当該原則は、①包摂的な成長、持続可能な開発及び福利（well-being）、②人間中心の価値・公正性、③透明性・説明可能性、④頑健性・セキュリティ・安全性、⑤説明責任の5項目からなる。

「成長」では、関係者は人々や地球に役立つ便益を追求すべきであるとし、例として「人間の能力の拡張」「人間の創造性の強化」「包摂性の推進」「不平等の解消」「自然環境の保護」に向けたAI活用を提言する。「価値」では、「法の支配・人権・民主的価値」を尊重すべきであり、それらには「自由」「尊厳と自律」「プライバシーとデータ保護」「無差別と平等性」「多様性」「公正性」「社会的正義」「労働権」が含まれるとする。また、「価値」の尊重においては「人間が自ら決定する」仕組みを導入すべきである

---

[46] G7, *"G7 People-Centered Action Plan on Innovation, Skills and Labor"*, 2017/05/27

[47] 共通ビジョンは①人間中心のAIの推進、②AI投資の促進、③生涯学習の支援、④AIアプリの開発から実装までの全ての段階で多様な人々の関与を確保、⑤ステークホルダーによる対話の促進、⑥AIの信頼を促進する努力を支援、⑦中小企業等によるAIアプリ利用の促進、⑧労働市場政策、労働者訓練の促進、⑨全ての人々にチャンスをもたらすAIへの投資の奨励、⑩デジタルセキュリティ改善の動きを奨励、⑪プライバシー・個人データ保護の確保、⑫差別的な貿易慣行（強制的な技術移転、不当なデータローカライゼーションなど）の排除という12項目にコミットすると表明する（G7, *"Charlevoix Common Vision for the Future of AI"*, 2018/06/09）。

[48] G7, *"G7 Hiroshima Leaders' Communiqué"*, 2023/05/20(para. 38)

[49] G7, *"Hiroshima Process s International Guiding Principles for Organizations Developing Advanced AI System"*, 2023/10/30

[50] OECD, *"Recommendation of the Council on Artificial Intelligence"*, 2019/05/22

とする。「透明性」では、透明性と責任ある情報開示を同列に並べた上で、一般的な知識の涵養、AI システムと交信していることの開示、出力に関する説明、異議手続の整備を勧告する。

3．UNESCO「人工知能倫理勧告」（2021 年 11 月）

UNESCO は科学倫理に関する諮問機関（COMEST）を設置し、「ロボット倫理に関する報告書[51]」（2017 年 9 月）、「AI 倫理に関する基礎的研究[52]」（2019 年 2 月）などを発表してきた。それらの成果を踏まえて 2021 年 11 月に「AI の倫理に関する勧告[53]」を採択した。そこでは全体の基礎となる「5 つの価値」とこれを行動に落とし込む「10 の原則」を説明する。

価値として①人権・基本的自由・人間の尊厳の尊重・保護・促進、②環境・生態系の繁栄、③多様性・包摂性の確保、④平和で正義のある繋がる社会に生きることを掲げる。「多様性・包摂性」では全員の積極的な参画とともに、AI システムを使う使わないの自由な選択など個人の尊重にも言及する。「繋がる社会：interconnected society」では、生物同士の繋がり、生物と環境との繋がりが平和で正義のある社会に住む価値を高めるという説明を付す。

原則として(a)比例原則と無害性、(b)安全とセキュリティ、(c)公正性と無差別、(d)持続可能性、(e)プライバシー権とデータ保護、(f)人間による監督と決定、(g)透明性と説明可能性、(h)対応責任と説明責任、(i)認知とリテラシー、(j)多様な関係者、適応性のあるガバナンスと協力を掲げる。「無害性」では不可逆的な結果をもたらす決定・人の生死に関する決定は人間が行うべきであるとする。「公正性」では生成物による差別の禁止だけでなく利用・アクセス面での公正性、デジタル格差の解消にも言及する。「人間による監督」と「対応責任」では責任主体の不在をなくすために「倫理的・法的責任を人間その他の法的主体に割り当てること」を提言する。また「説明責任」の中で「監査可能性」「追跡可能性」に言及する。

4．IDCPPC「人工知能開発に関する指導原則」（2018 年 10 月）

各国データ保護当局の連合体である「データ保護・プライバシー委員会国際会議：IDCPPC」は 2018 年 10 月に採択した「人工知能における倫理・データ保護に関する宣言[54]」の中で「人工知能開発に関する指導原則」を提言する。

指導原則は①基本的人権の尊重と公正の原則、②注意、警戒と説明責任、③透明性と理解可能性、④プライバシーの保護、⑤能力の向上・権利の実行・大衆参加の確保、⑥不法なバイアス・差別の排除という 6 項目からなる。

「人権」では当初設定した目的にそって AI もデータも利用されること、AI 利用の影響を個人単位でも集団単位でも考慮すること、人間の成長を阻害し危険に晒すように開発されないことに言及する。「注意・警戒」では監査・モニタリング・影響度評価・定期検証を通じて説明責任を果たすこと、基準の開発や好事例の共有による集団的・共同的な責任を涵養すること、AI に関する知識や社会的影響に関する教育・研究に投資すること、民主的なガバナンスプロセスを確立することに言及する。「透明性」では説

[51] COMEST, *"Report of COMEST on Robotics Ethics"*, 2017/09/14
[52] COMEST, *"Preliminary Study on the Ethics of Artificial Intelligence"*, 2019/02/26
[53] UNESCO, *"Recommendation on the Ethics of Artificial Intelligence"*, 2021/11/23
[54] IDCPPC, *"Declaration on Ethics and Data Protection in AI"*, 2018/10/23

明可能なAIに関する研究の推進、透明性・理解可能性・リーチ可能性の促進、組織の実務の透明化、情報に関する自己決定権の保証、AIの目的と影響に関する情報の提供に言及する。「能力向上」ではデータ保護・プライバシー権の尊重、表現の自由など関連する権利の尊重、出力結果に反対する権利や異議申出の権利の確保、AIを使った平等な能力向上や大衆参加の実現に言及する。

５．GPA「人工知能システムに関するリスク管理フレームワーク」（2022年10月）

IDCPPCの後継組織である「国際プライバシー会議：GPA」は2022年10月に作業部会報告書「AIシステムの一般的リスク管理枠組み[55]」を発表した。

リスクに晒される「価値」として①公正性と合法性、②透明性と説明可能性、③救済の申立と実行、④データの最小利用と保管期間の制限、⑤利用目的による制限、⑥データの中身・質の正確性、⑦説明責任と賠償責任、⑧データセキュリティ、⑨倫理的配慮を掲げる。

「公正性」ではデータ処理・出力における公正性に加えて、人の脆弱性の悪用、不平等・差別・社会的分断の悪化、不法な物理的・非物理的加害のために利用しないことを求める。「透明性」では情報の取扱いに関する自己決定権（informational self-determination）を掲げ、そのためにAIとの交信や個人データの提供に際して必要な情報を提供すべきだとする。また、個人に関する推論・予測・プロファイリング・分類の実施や商品・サービスの推奨に用いられる場合にはデータアクセス権の確保が重要であるとする。「責任」では負の影響に関する責任（responsibility）、関係者に対する説明責任（accountability）、損害を補償する賠償責任（liability）を説明する。「倫理的配慮」ではAI利用の必要性とその負の影響のバランスをとる必要があるという立場を明らかにし、社会的スコアリングや個人の自動識別に用いられる場合には法的枠組みを遵守すること、ビッグデータを用いた統計的推論を行う場合にはバイアス・差別に留意すること、開発者倫理が重要であることに言及する。開発者倫理では特に「無害性（セキュリティと肉体的・精神的安全性）」「有益性（福利・人間の尊厳・平等性・持続可能性・連帯）」「出力の正義と公正性」「賠償責任」「内部通報の保護」「自律性（自己決定・選択の自由）」に言及する。

管理すべき「リスク」として(a)倫理原則違反、(b)透明性の欠如、(c)個人データ保護原則違反、(d)不公正な差別、(e)個人データの不法・不公正な支配、(f)人の脆弱性の悪用、(g)個人に対する加害、(h)社会に対する悪影響（偽情報・ディープフェイクなど）、(i)持続可能性への負の影響を挙げる。

６．小括

国際機関における5つの「AI原則」を比較してみると（《図表5》）、「個人的課題」「社会的課題」「技術的課題」に万遍なく言及しており、欧州での検討の影響が色濃い。

[55] GPA AIWG, *"Risks for Rights and Freedoms of Individuals Posed by AI Systems: Proposal for a General Risk Management Framework"*, 2022/10/25

《図表 5》国際機関の AI 原則

区分	項目	G7 2023/10	OECD 2019/05	UNESCO 2021/11	IDCPPC 2018/10	GPA 2022/10
個人	人間の尊厳		△	○		△
	人権の尊重	○	△	○	○	
	プライバシー	△	△	○	○	○
	公正性・公平性	○	○	○	○	○
	平等（無差別）	○	△	○	○	△
	自律性・自己決定権		△	○	△	△
社会	民主主義	○	△			
	法の支配	○	△			
	包摂性		○	○	△	
	多様性	○	△	○		
	広義の責任			○	△	△
	賠償責任			△		○
技術	安全性	△	○	○		
	セキュリティ	○	○	○		○
	無害性・危害防止	△		○		△
	透明性	○	○	○	○	○
	説明責任	○	○	○	○	○
	説明可能性		○	○		○

（注 1）比較する原則は左から「G7 人工知能開発組織向け指導指針」「OECD 信頼できる人工知能のための責任あるスチュワードシップ原則」「UNESCO 人工知能倫理勧告」「IDCPPC 人工知能開発に関する指導原則」「GPA 人工知能システムに関するリスク管理フレームワーク」である

（注 2）「○」は見出しに使われる概念を、「△」は説明文の中に登場する概念を表す

（出典）当社作成

Ⅶ. AI 原則で使われる概念の整理

各国政府・機関は自主規制（関係者に遵守を求める）である「AI 原則」をそれぞれに提言しており、そこには多数かつ多様な概念が盛り込まれている。その数は、見出しから取り出しただけでも 60 余りに上る。幅広い関係者と多彩な利用事例（use case）を網羅的にカバーしようとしたためである。その点を理解せずに、これらの概念を横一直線に並べても、何かを導き出すことは難しい。自分がどの立場にあり、どの利用事例に該当するかにより、使える概念と使えない概念が出てくる。そこで、体系的な理解や取捨選択に資する分類方法をいくつか考えてみる。

１．守るべき「価値」ととるべき「行動」

多くの AI 原則は、プライバシーの保護や説明責任の履行など、いくつかの概念を同列に並べている。その中で、欧州委員会「倫理ガイドライン」（Ⅳ－３）は「3 つの要素」「5 つの基本権」「4 つの倫理原則」「7 つの要件」という多重構造を採用し、UNESCO「倫理勧告」（Ⅵ－３）も「価値」と「原則」という二層構造になっている。

守るべき「価値」ととるべき「行動」に敢えて分けるとすれば、価値として「人間の尊厳」「人権の尊重」「民主主義」「法の支配」などを挙げることができる。企業がより実践的な枠組みを構築するのであれば、「行動」を中心に整理することが考えられる。

２．誰の課題か（個人・社会・技術）

人工知能のもたらすデメリット（課題）を誰が被るのか（対象者）という観点から「個人にとっての課題」「社会にとっての課題」「技術上の課題」という三分法が出てくる。なお、技術上の課題に対処すべき名宛人は、設計・開発者に絞られる。また、「公正性」や「説明責任」のように複数の区分に当てはまるものも存在し、排他的な区分という訳でない。

（１）個人にとっての課題

「個人にとっての課題」は、①不当な取扱いを受けること、②正当な権利を侵されること、③自らの意思を歪められることに分けることができる。

①不当な取扱い

人々を分類するツールとして人工知能を使う場合（融資審査など）に、「バイアス」のあるデータを学習した結果、特定の人を「差別」することが想定される。この関係で「人間の尊厳」「法の下の平等」「公正性」などの概念が使われる。また「無差別」であるとの安心感に向けて「透明性」「説明責任」「監査可能性」などが併用される。

②権利の侵害

何らかの生成物（文章・画像など）を得る目的で人工知能を使う場合に、学習データとして取り込んだもの（プライバシー情報、著作物など）をそのままの形で生成することが想定される。また、特定の個人（犯罪者など）を追跡するツールとして人工知能を使う場合に「監視社会」に対する警戒が併せて議論される。これらの関係で「人権の尊重」「プライバシー」「個人データの保護」「知的財産権の保護（合法性）」などの概念が使われる。

③自由意志の侵害

人々の嗜好を特定するツールとして人工知能を使う場合（ターゲティング広告など）に、他の選択肢が示されないことから「欲しくない商品を買ってしまった」「過激な意見に感化されてしまった」という事態が生じ得る。逆に、人工知能による推奨への依存が進み「人工知能なしでは生活ができない（AI 依存）」という事態も想定される。これらの関係で「自律」「自由」「自己決定権」などの概念が使われる。

《コラム 4》人工知能と自己決定権

人間の「自律性」を尊重すべきであるという原則に言及する事例は多い。この「自律性」については、自分のことを自分で決定すること（自己決定権）と、AI により自らの意見・感情・行動を操作されないこと（AI からの自由）の両面で説明される。このうち「自己決定権」では、①AI システムを利用するか否かを決定すること（米国権利章典）、②AI システム利用中にあっても自ら決定すること（欧州倫理ガイドライン）、③自らの個人情報の取扱いを自ら決定すること（GPA リスク管理）といった具体化がなされている。また、「自由」では、①機械的に出された決定に従わない権利（欧州倫理ガイドライン）、②プロファイリングを受けない権利（欧州倫理 G 声明）、③人間の分類・採点・操作を目的に AI を開発しない義務（欧州倫理ガイドライン）といった権利・義務に言及する例がある。

（２）社会にとっての課題

「社会にとっての課題」は、①社会秩序が混乱すること、②排除される人々が生まれること、③世の中が受け入れないことに分けることができる。なお、①と②は、SNS が社会に与える課題と共通する。

①社会秩序の混乱

生成 AI は、「確からしさ」をまとった生成物を大量に生成する能力があり、その悪用を考える人（悪意者）と生成物を拡散させる道具（SNS など）が揃うと、社会秩序の混乱（世論操作など）を招くことができる。この関係で「（個人でなく）人類の福利（well-being）」「民主主義」などの概念が使われ、また、「頑健性」概念の中で悪意利用の排除に言及する。

②一部の人々の排除

「排除」として(a)AI 関連の商品・サービスにアクセスできない人が生じる（デジタル・デバイド）、(b)AI による生産性の向上により仕事を奪われる人が出るという事態が想定されている。これらに対処すべく「公正性」「正義」「包摂性」という概念が使われる。また、「排除」を生まないルール作りにおいて、多様な人々の意見を反映することが重要であり、「多様性」という概念も使われる。

③世の中が受け入れない

産業革命期のラッダイト運動[56]のように、「過度な懸念」を持つ人々が人工知能の受入れを拒絶する事態が想定される。これを避けるために提唱されるのが「信頼できる人工知能」であり、積極的な情報発信により「過剰な懸念」を取り除くという文脈で「透明性」と「説明責任」という概念が使われる。なお、「説明責任」は、世の中一般の理解を深めるための「説明」と、個別の利用者の照会に対応する「説明」の 2 種類があり、後者は「個人にとっての課題」に対応するものとなる。

また、人工知能に起因する事故に備えておくことも「世の中の受入れ」に必要である。この関係で「賠償責任 liability」という概念が使われている。こちらも、実際の被害者から見れば「個人にとっての課題」となる。

《コラム 5》人工知能をめぐる責任

人工知能が何らかの問題を起こしたとしても、①様々な形で社会に実装され（開発者だけでなく幅広い人々が実装までに関与している）、②実装後も進化する（実装後も学習を継続しており、どの段階の学習が問題に繋がったか特定できない）といった特徴があるために、責任関係を明確にすることが難しい。そこで「責任のある人を特定する」アプローチでなく「関係者の間で責任を割り振る」アプローチをとる AI 原則は多い（欧州倫理 G 声明）。その「責任」にあっても、英語文献では複数の単語（responsibility・accountability・liability）が使い分けられている。

第一の「responsibility」は、各原則において「設計・開発・調達・利用における人間の役割と責任

[56] 産業革命期の織物工業界において、織機の機械化が熟練者の失業と非熟練者の労働強化をもたらしたと考えた労働者は、1810 年代に機械や工場設備を相次いで打ち壊した。

を定義し、理解させ、割り当てるべきである」（米・政府利用原則）、「責任の原則から言えば自律システムは地球規模の社会的・環境的な福利に奉仕するよう開発・利用されるべきである」（欧州倫理 G 声明）、「法的な問題に対処するためには責任の配分を決める解決法と法令に拘束力を持たせる仕組みに投資すべきである」（同）という使われ方をしている。利用者の照会に答える、社会的責任を果たすといったかなり広い概念である。第二の「accountability」は「説明責任」と訳され、一般的な研究倫理（開発者が自ら開発した技術について説明責任を負う）だけでなく、医療分野における「インフォームドコンセント」や金融分野における「商品説明義務」など、事業者が具体的な責任を負う局面も想定される。「情報提供や説明を拒否する理由として企業秘密を持ち出すべきでない」（GPA）という指摘もある。なお、ブラックボックス問題の関係では「explainability（AI の機能や出力に関する説明可能性）」という用語が用いられている。第三の「liability」は発生した損害の補償（賠償責任）という文脈で用いられることが多い。例えば、「自律システムの振る舞いにより引き起こされた損害に対して誰が責任を負うのかを明確にすべきである」（欧州倫理 G 声明）といった記述がある。

（３）技術上の課題

「技術上の課題」は、①AI を搭載した商品・サービスが利用者に危害を加えること、②AI の生成物について誰も説明できないことに分けることができる。なお、利用者への加害は、人工知能に特有の課題ではなく「製品・サービス」一般に当てはまる課題である。

①人体・財物への加害

ロボット・自動走行車などの AI 搭載製品では、誤作動やセンサー異常により人体・財産を損なう事態が想定される。この関係で「安全性」「正確性」「信頼性」「無害性」などの概念が使われる。搭載製品の種類によっては安全性を確保する既存法制[57]があり、「法令遵守」「法令適合性」という概念でカバーすることもある。また、通信回線を利用する AI 搭載製品・サービスにあってはサイバー攻撃による誤作動の誘発も想定される。この関係で「セキュリティ」「頑健性」という概念が使われる。結果に対する責任として「賠償責任」も問題となる。

②出力結果の説明

人工知能は、機械学習を通じて確からしい「手順」を発見し、自ら編み出した「手順」にそって生成物を生み出す。そのプロセスに「人間」は介在しておらず、開発者であっても「手順」を説明できない。ところが、AI 搭載製品・サービスの提供者・利用者は、最終利用者に対して「手順」を説明する義務を負う。この関係で「説明責任」「説明可能性」「理解可能性」「透明性」という概念が使われる。

③課題の技術的な解決

「個人にとっての課題」「社会にとっての課題」「技術上の課題」に共通して、仕組みで解決する概念も使われる。「人間による監督」「制御可能性」「監査可能性」「結果への配慮」などである。

[57] 道路運送車両法（自動車）、航空法（ドローン）、医薬品医療機器法（医療機器）、消費生活用製品安全法（家電）など。

《図表 6》個人・社会・技術による分類

区分	懸念される事態	対応する概念
個人	不当な取扱いを受けること	人間の尊厳、法の下の平等、公正性、透明性
	正当な権利を侵されること	人権の尊重、プライバシー、個人データ保護、合法
	自らの意思を歪められること	自律、自由、自己決定権
社会	社会秩序が混乱すること	人類の福利、民主主義、頑健
	排除される人々が生まれること	公正、正義、包摂、多様
	世の中が受け入れないこと	信頼、透明、説明責任、賠償責任
技術	利用者に危害を加えること	安全、無害、合法、頑健、セキュリティ、賠償責任
	生成物を誰も説明できないこと	透明、説明責任、説明可能、理解可能
	課題を技術的に解決すること	人間による監督、制御可能、監査可能、結果への配慮

（出典）当社作成

《コラム 6》ブラックボックス問題と「説明可能な人工知能」

人工知能の「自律性」は、その出力結果を人間が説明できない（ブラックボックス）という状況をもたらす。ところが、公的給付の支給決定に人工知能を使う場合など「説明できない」では済まされない局面も存在する。そこで「説明可能な人工知能（XAI: explainable AI）」を目指す研究が進められている。米・国立標準技術研究所は、XAI の満たすべき要件として、①すべき説明（全ての出力に証拠と理由を付ける）、②意味ある説明（個人利用者に理解できる）、③正確な説明（出力を生成するプロセスを正しく反映する）、④運用の限界（設計時に想定した条件下又は出力に信頼を置ける状況下になければ AI を運用しない）という 4 項目を掲げる 。具体的な技術動向については亀谷由隆「説明可能 AI 技術のこれまでとこれから」（電子情報通信学会レビュー、2021/10）などが参考になる。

３．いつからの課題か（既存・新規）

人工知能については、上記２記載のとおり、様々な問題事象を想定した上で、それを解決する拠り所（原則）として様々な概念が使われる。しかし、想定する問題事象の全てが「人工知能に特有の課題（新たな課題）」であるとは言えない。既存の課題にあっては、既存の議論や枠組みを応用することを先ず検討すべきである。

（１）既存の課題

上記２における「不当な取扱い」（（１）①）や「権利の侵害」（（１）②）は、人間が生成物を作り出す場合においても生じる課題であり、人間の「判断ミス」が人工知能の「判断ミス」に置き換わったに過ぎない。また、「自由意志の侵害」（（１）③）や「社会秩序の混乱」（（２）①）や「一部の人々の排除」（（２）②）は、異なる切り口、すなわち媒体となる SNS などがもたらす社会的課題として既に議論されている。これらの課題にあっては、人工知能が「問題」を増幅する可能性はあるものの「新たな論点」が加わる訳でない。更に、「人体・財物への加害」（（３）①）は、AI 非搭載の製品において既に問題とされている。ただ、「責任主体の不在」などの新たな要因が加わるため、「責任のあり方」などの論点が追加されることになる。

（２）新規の課題

こうして見ると、人工知能のもたらす「新たな課題」と言い得るのは、ブラックボックス問題に起因する「出力結果の説明」（（３）②）とその帰結である「責任主体の不在」（（２）③）の 2 点に止まる。

各国政府の AI 原則は様々な概念を掲げるが、企業（開発企業・利用企業）は、リスクを洗い出す段階で「既存の課題／新規の課題」に仕分けを行い、前者にあっては既存の枠組みの見直しを、後者にあっては新たな枠組みの構築を検討するのが効率的であろう。

４．小括

「価値／行動」の二分法は、より実践的な「自主規制」に取り組む際に有意義である。ただ、求められる複数の行動がお互いに矛盾する事態も想定され、将来的には、両者を調整する「価値」を持ち込むことも念頭に置くべきであろう。「個人／社会／技術」の三分法は、自社の立ち位置や利用事例に基づいて「概念」を取捨選択する際に有意義である。単に AI を利用するに過ぎない企業は「技術」に対応する必要はなく、AI 製品・サービスを提供する企業は「社会／技術」に配慮する必要が出てくる。「既存／新規」の二分法は、既存の「自主規制」の中に転用可能な枠組みがあり得ることを示唆する。いくつかの分類方法を併用すれば、自社に相応しい「概念の取捨選択」に資するものと思われる。

《コラム 7》人工知能の技術としての特異性

生成 AI の登場とともに高まる「懸念」は、他の革新的な技術に見られないものとなっている。技術面で見た「特異性」について、各国 AI 原則[58]の中から抽出してみる。

１．人の手から離れる（自律）

人工知能は、人間の「脳」に代わって知的活動（文章・画像の生成など）や身体活動（自動車の操縦など）を行うべく開発されてきた（代替性）。ところが、「完全代替」の未来が見通せるようになると、人々は「排除」される懸念を抱くようになる。代替のために開発された人工知能が、その「代替性」の故に警戒されている。従来は、人間が「人工知能の動き方」をプログラムという形で定義してきた。その段階の動き方は「設計者が想定した範囲」に止まり、安全性などを事前に「検証」することも可能であった。ところが、「機械学習」が導入され、学習が人の手を離れると、最終的な成果物（文章、機器制御など）の範囲も精度も、人の手を離れて拡大・向上する（自律性）。ここに「完全代替」の可能性が生まれ、人間の「懸念」も抽象的なものから現実的なものへと転じることになる。

２．自ら進化を続ける（進化）

「AI の自律性」が製品・サービスの出荷・提供開始段階で固まるのであれば、「完全代替」の可能性をコントロールすることができる。ところが、人工知能はその後も「進化」を続ける。製品・サービスを利用する段階でも、人々による指図（プロンプト）を学習データとして用いるからである。この点は、「完全代替」の懸念を増加させるだけでなく、懸念に対応すべき主体（責任主体）を拡大する。従来の技術であれば、責任主体は開発者や販売者に止まっていたのに対し、人工知能にあっては製品・サービスの利用者まで「進化に伴う課題」に対処する責任を負うことになる（《コラム 1》参照）。

[58] 米国国立標準技術研究所「人工知能リスク管理フレームワーク」は、人工知能の「特異性」として①学習データが時々刻々と変化して人工知能が機能面で影響を受ける、②システムや利用環境が様々で失敗の検知が難しい、③社会全体の動きや人間の振る舞いなどの社会的要素もリスクに影響するという点を挙げる（同書１頁）。

３．何にでも使える（多用途）

人工知能の能力は、現時点では人間の「脳」に比べて著しく劣っており[59]、「完全代替」（汎用人工知能 AGI: Artificial General Intelligence）の見通しが立っている訳でない。それでも、専門家の知見を初心者に教えるシステム、公的給付や融資の可否を判断するシステム、自動車・ドローンを操縦するシステムなど幅広い利用領域（use case）が生まれており、「人間の領域を徐々に侵している」感覚を生み出し続けている。

こうした「懸念の蓄積」は、何らかの出来事（インシデント）を契機として「炎上」する可能性が高い。企業にとっては「制御し難い」リスクである。

Ⅷ．おわりに

AI に対する注目は、期待と懸念の両面で極めて高まっている。各国・機関の AI 原則は、「懸念」に対応するものであるが、遵守を求める相手方の多様性（開発者から利用者まで）や利用事例の広がり（文章生成から機械制御まで）を反映して、極めて多岐にわたる「懸念」が想定されている。

一方で、企業は、AI を搭載した製品・サービスを利用するに過ぎないとしても、自主規制（既存のガバナンス・リスク管理体制への組み込み）に取り組む必要がある。その際に、各国・機関の AI 原則が想定する「全ての懸念」に対処する必要はなく、自らの立ち位置（製品・サービスの提供、自社向けカスタマイズ、最終利用など）や利用事例（文章生成、音声対応、機器制御など）に即して「懸念」を取捨選択し、そこで想定される「不測の事態」に備えていくことが効率的である。第Ⅶ章で説明した分類方法は、この取捨選択に役立つものと思われる。

＜参考文献＞

・AI ネットワーク社会推進会議報告書（2016/10、2017/07、2018/07、2019/08、2020/07、2021/08、2022/07）
・AI 白書編纂委員会「AI 白書 2023」（KADOKAWA、2023/05）
・科学技術振興機構「人工知能研究の新潮流２」（2023/07）
・総務省「情報通信白書」（平成 28 年版、平成 30 年版、令和元年版）
・Ad Hoc Committee on AI, *“Feasibility study on a legal framework on AI design, development and application”*, 2020/12
・EU Joint Research Centre, *“AI: a European perspective”*, 2018/12
・Freedom Online Coalition, *“FOC Joint Statement on AI and Human Rights”*, 2020/11
・Future of Life Institute, *“Asilomar AI principles”*, 2017/08
・IEEE (US), *“Ethically aligned design: 1st edition”*, 2019/03
・WHO, *“Ethics and governance of AI for health”*, 2021/06

[59] 人工知能のパラメータ数（ChatGPT で 1,750 億）と人間の脳のシナプス数（100 兆以上）を比較することが多い。