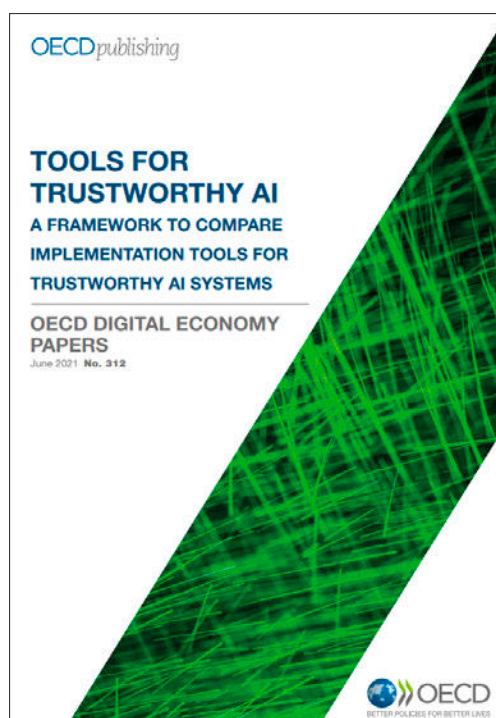


第4節 人間とAIの健全な協働に向けて

OECDは、「信頼できるAI」について、「OECDのAI原則を体現するAIシステム、すなわち、人権及びプライバシーを尊重し、公平性、透明性、説明可能性、頑健性、安全性およびセキュリティを備え、かつ、その開発および利用に関与する主体が説明責任を負うAIシステム」を指すとしている（図表 I-3-4-1）^{*1}。

図表 I-3-4-1 OECD “TOOLS FOR TRUSTWORTHY AI”



（出典）OECD（2021）^{*2}

第3章

AIの進展に伴う課題と対応（人間とAIの健全な協働に向けて）

日本においても、本稿「はじめに」で記載した「8項目の原則」が2016年に示された後、2019年に「人間中心のAI社会原則」にて、AIを人類の公共財として活用し、社会の在り方の質的变化及び真のイノベーションを通じて地球規模の持続可能性とつなげることが重要であるとされるとともに、「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）」・「持続可能な社会（Sustainability）」の3つの価値が「基本理念」として示され、「その実現を追求する社会を構築していくべき」とされた。

現在、これらの基本理念の実現については、「AI事業者ガイドライン」において各主体が念頭に置くべき原則が「共通の指針」として整理されている。具体的には、各主体が取り組む事項として1) 人間中心、2) 安全性、3) 公平性、4) プライバシー保護、5) セキュリティ確保、6) 透明性、7) アカウンタビリティの7原則、そして、社会と連携した取組が期待される事項として8) 教育・リテラシー、9) 公正競争確保、10) イノベーションの3原則、合計10の原則が示されており、例えば、1) 人間中心においては「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

^{*1} OECD（2021）“TOOLS FOR TRUSTWORTHY AI：A FRAMEWORK TO COMPARE IMPLEMENTATION TOOLS FOR TRUSTWORTHY AI SYSTEMS”（OECD DIGITAL ECONOMY PAPERS June 2021 No. 312）〈https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/06/tools-for-trustworthy-ai_0e36bb08/008232ec-en.pdf〉（2026年4月7日参照。訳文は総務省作成。）

^{*2} 同上。

図表 I-3-4-2 各主体に共通の指針・広島AIプロセス「全てのAI関係者向けの広島プロセス国際指針」
(AI事業者ガイドライン)

本編	別添	各主体に共通の指針【1/2】	本編	別添	各主体に共通の指針【2/2】
第2部		・各主体が連携し/リユース全体で取り組むべき「共通の指針」は、以下のように整理されます ・なおこれらの取組は、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自主的に進めることが重要です	第2部		・各主体が連携し/リユース全体で取り組むべき「共通の指針」は、以下のように整理されます ・なおこれらの取組は、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自主的に進めることが重要です
指針		内容（主な項目の抜粋）	指針		内容（主な項目の抜粋）
1) 人間中心		- AI だけでなく多様な幸せ、多様な人々の多様な幸せ (well-being) の追求が可能なよう行動する - AI が生成した価値・価値観・倫理観が社会や文化に適合しない場合、適正なリスク管理を確保することを経営上として必要対策を講ずる - よその人々がAIの恩恵を享受できるように社会的弱者によるAIの活用を容易にするよう注意を払う	各主体が取り組む事項		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開いた適切な情報で適切な情報を提供する (AIを用いているという事実、活用している組織、データ交換及びデータフローの手法、AIシステム・サービスの能力、限界、提供先における適切な不適切な利用方法、等)
2) 安全性		- 適切なリスク分析を実施し、AIシステムの評価を適切に - 主体のAIシステム及び組織で本来の用途以外に提供・利用により危害が発生することに対し - AIシステム・サービスの特性及び用途と照らし、学習等に用いるデータの正確性を検証するとともに、データの透明性の確保、適切な仕組みの遵守、AIモデルの更新等を合理的な範囲で適切に実施する	6) 透明性		- 各主体のAIがAIシステムに関するポリシー・サービス・ガイドライン等の方針を策定し、公表する - 関係する関係者に提供し一定期間保管し、必要なところで、入手可能な範囲に限り適切に公開し利用方法、等)
3) 公平性		- 特定の個人ないし集団への人種、性別、国籍、年齢、政治的意見、宗教等の多様な背景を理由として不当で有害な扱い及び差別をなすよう努める - AIの出力結果が公平性を欠くことのないよう、AIに単独で判断をさせるだけでなく、適切なチームで人間の判断を介しながら利用を検討し、潜在的なバイアス等を把握し、AIの刷新・適切な利用を行う	7) アカウンタビリティ		- トランパシリティの確保と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がAIシステムに関するポリシー・サービス・ガイドライン等の方針を策定し、公表する - 関係する関係者に提供し一定期間保管し、必要なところで、入手可能な範囲に限り適切に公開し利用方法、等)
4) プライバシー保護		個人情報保護法その他の関連法令の遵守、各主体のプライバシーポリシーの策定・公表により、社会や文化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、保護されること。その重要性に応じた対応を取る	8) 教育・リテラシー		- AIに関わる者らが、その役割において十分なAIリテラシーを確保するために必要な措置を講ずる - AIの特性、情報に依って特性及び用途的な適用の可能性もあることを勘案して、ステークホルダーに依って教育を行うことが期待される
5) セキュリティ確保		- AI システム・サービスの信頼性・完全性・可用性を確保し、特に、AIの安全性を高め、その時点での技術的・制度的に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、保護されること。その重要性に応じた対応を取る - AIシステム・サービスのセキュリティに関する手法だけでなく、このリスクに対するAIシステムの価値と価値を確保する	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を促す公正な競争環境の構築を促すことが期待される
10) インベリション		- 国際化・多様化、デジタル変革及びオープンイノベーションを推進する - 各主体のAIシステム・サービスの相互接続性及び相互運用性を確保する - 標準化がある場合には、それに準拠する	10) インベリション		- 国際化・多様化、デジタル変革及びオープンイノベーションを推進する - 各主体のAIシステム・サービスの相互接続性及び相互運用性を確保する - 標準化がある場合には、それに準拠する

(出典) 総務省・経済産業省「AI事業者ガイドライン(第1.2版)」概要^{*3}

AI技術自体が刻一刻と今も進化を続け、また、社会においてAIが適用される分野・領域も拡大していることから、「信頼できるAI」の具体的な在り方については、AI基本計画がその原則の一つに定める「アジャイルな対応」に基づき、今後も各主体における不断の検討・熟考・議論の下に実現されていくことが重要であると考えられる。また、AIが技術の一つである以上、その機能を効用として享受できるか、損失や損害として被ることになるかは、AIを開発・利用する我々人間の側にかかっており、こうした意味において、「信頼できるAI」と表裏をなすのは、AIを過信も不信もせず、適切な距離感のもとにAIを信頼できる人間の理性的な力であるとも考えられる。

「信頼できるAI」の追求とともに、このような「適切にAIを信頼できる力」の涵養も不断に追求していくことも重要であり、人間とAIが健全に協働するデジタル社会は、こうした「信頼できるAI」と「適切にAIを信頼できる力」の双方を目指した先の地平に、広がっているであろう。

*3 総務省・経済産業省（2026）「AI事業者ガイドライン（第1.2版）」概要
 〈https://www.soumu.go.jp/main_content/001064299.pdf〉（2026年5月12日参照）