［投稿論文：総説・レビュー論文］

AI 原則の国際的な潮流と日本の AI 原則の特質

今後の見直し検討に向けた政策上の示唆

Idiosyncrasies of the Japanese AI Principles in Relation to International Issues and Principles

Policy Implications for Possible Future Review of Them

畠山 記美江
慶應義塾大学大学院政策・メディア研究科特任助教（有期）

Kimie Hatakeyama
Project Research Associate (Non-tenured), Graduate School of Media and Governance, Keio University

Correspondence to: kimie@sfc.keio.ac.jp

大磯 一
慶應義塾大学 SFC 研究所上席所員

Hajime Oiso
Senior Researcher, Keio Research Institute at SFC

Abstract:　AI に関連する新たな問題への対応のため官民を問わず国内外の様々な団体から公表されてきた非拘束的な指針（いわゆる AI 原則）については、各原則中で言及される項目の種類に一応の合意があるとする見解と、見かけ上の合意に過ぎないとの見解の両方が存在する。本研究では主に、総務省の AI 原則について欧州評議会の人工知能特別委員会の分析結果を用いて国際比較を行い、同一に見える項目にも、安全保障を含む緊急事態対応や持続可能性等、国際的な議論との差異があることを示す。今後の政策検討では、日本独自の点の訴求と、差異への対応の双方が課題となり得る。

Various public and private organizations have published so called ‘AI principles’ i.e. non-binding guidelines to address new issues relevant to AI. There are both views that rough consensus exists across principles on the types of their items and that such consensus is only superficial. This study mainly makes an international comparison on the principles issued by MIC of the government of Japan using the analysis released by CAHAI of Council of Europe, and shows that even items that appear to be identical, some of them, such as emergency preparedness including national security and sustainability, have differences in contents between Japan and the rest of the world. Addressing the differences while disseminating Japan’s uniqueness could be a future issue.

Keywords: AI 原則、AI 倫理、AI 政策
AI principles, AI ethics, AI policy

1　はじめに

1.1　背景

生成 AI、顔認証、自動翻訳やターゲティング広告、チャットボット、運転支援、品質検査、感染症拡大予測、医療における画像診断支援、投資、教育、婚活にいたるまで、AI は日常生活の様々な場面に組み込まれており、もはや人間の社会活動とは切り離し得なくなっている。そして、AI はこれまでの価値観だけでは対処しきれない問題を数多く浮かび上がらせた。

中川 (2020) によれば、2045 年には AI が人間の能力を凌駕する、というシンギュラリティ論 (カーツワイル , 2007) に端を発した AI 脅威論から人間が AI を制御するための方法として AI 倫理の議論が活発になり、AI を安全に実装・運用するための非拘束的な指針として、様々な団体が AI 原則を公表するようになった。日本政府は AI 原則策定の草創期から検討に着手し、2017 年に「国際的な議論のための AI 開発ガイドライン案」(総務省 , 2017) (以下、「開発ガイドライン」)、2019 年に「AI 利活用ガイドライン」(総務省 , 2019b) (以下、「利活用ガイドライン」) が総務省から公表され、次いで、内閣府から「人間中心の AI 社会原則」(内閣府 , 2019) が公表された。欧州評議会の人工知能特別委員会 (CAHAI) から 2020 年 7 月に公表された報告書「AI 倫理ガイドライン 欧州と世界の視点ーマルセロ・イエンカ、エフィ・ヴァイエナによる暫定報告書」(CAHAI, 2020) (以下、「CAHAI 報告書」) によれば、AI 原則の公表数は 2018 年にピークに達した (CAHAI, 2020, p.10)。2019 年 8 月 9 日に総務省 AI ネットワーク社会推進会議が公表した「報告書 2019 概要」において、「AI 原則の項目については、国際的にほぼコンセンサスが得られつつあり、今後は原則の実効性を確保するための具体的手段についての議論に移行。これらの議論に貢献し、認識の共有を図る」(総務省 , 2019a, p.3) との報告がある。加えて、2019 年 5 月 22 日に OECD において AI に関する初の国際的合意による原則、「人工知能に関する理事会勧告」(OECD, 2019) が採択されたことも、AI 原則について一定の合意がなされたという見解を強化する材料となってい

る。

確かに原則の項目だけを見ると、国際的に重要視される点が定まったように見受けられる。本論文では、総務省の 2 種類の AI 原則「開発ガイドライン」と「利活用ガイドライン」を題材に、これらの日本の AI 原則が国際的な議論の潮流の中でどのような状況にあるかを見極め、今後の政策検討に資する示唆を得ることを目指して実施した研究の結果を報告する。

2　研究の範囲

2.1　先行研究の概観

AI 原則の策定状況については、Hagendorff (2020)、新保 (2020)、福岡 (2020) など国内外の研究者が AI 原則における採用項目の一覧表を作成しており、中川 (2020) は AI 倫理指針で重要視されている項目を時系列でトレンドの分析をしている。公的機関による調査では、欧州評議会の人工知能特別委員会 (CAHAI) から 2020 年 7 月に報告された報告書 (CAHAI, 2020) があり、(1) 変化し続ける非強制的なガバナンス手段をモニターすること (2) AI が倫理原則、人権、法の支配、民主主義に与える影響を前向きに評価すること、を目的とし、機械学習を含む分析手法 (Jobin et al., 2019) を用いて、特に政府機関、非政府組織、学術機関、民間企業によって公表された合計 116 のガイドラインをレビューした結果を示す。日本では有識者会合「AI ネットワーク社会推進会議」[2] により 2022 年 7 月に公表された「報告書 2022」(総務省 , 2022) があり、国内外の主要な原則の調査、及びそれらと総務省の 2 原則との比較が行われている。本研究ではまず、公的機関による CAHAI 報告書と報告書 2022 の調査を紹介する。

2.2　CAHAI 報告書の概要

CAHAI 報告書の分析は、(1) 分析対象とする文書を特定するスクリーニングと、(2) 対象となった文書のコンテンツに対する量的及び質的な分析、の 2 段階から構成される。

第一段階のスクリーニングでは、PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) と呼ばれるフレームワークを応用し

た独自開発プロトコルにより、欧州連合基本権庁の AI 政策イニシアティブリストなど 4 つの情報源に対して検索が行われた。その上で、検索結果に含まれる全文書について一定の基準（CAHAI, 2020, p.7）により適格性が評価され、適格とされたものが分析対象とされた。

第二段階の量的分析では、まず、対象となった全ての文書群を機器の種類、発行者、発行者の地理的な出身地によって分類し、著者らが以前に開発した分析プロトコルの拡張版を用いた評価が行われた。その後、キーワード検索による全文書全文のスクリーニングが行われ、人権に言及している文書群である AI アプリケーションを設計、開発、展開する際に人権の保護と促進または侵害の防止のいずれかに明確に言及している文書群を特定し、その他の文書群との区別が行われた。

第二段階のうち質的分析としては、(1) 倫理原則と価値観、及び (2) 人権の各領域に関連して繰り返し現れる主題のパターンを特定する手作業が行われた。浮かび上がった主題パターンは、Jobin, Ienca & Vayena (2019) が開発した倫理マトリックスに基づき分析され、コード化され、あらかじめ定義された倫理カテゴリーに分類された。

この手動による主題分析を補完するものとして、自然言語処理（NLP）による自動分析も実施された。そこでは質的分析プロトコルの自動化された方式と手動の方式を併用しつつ、当該プロトコルで得られたコードから正規表現の構築が行われた。（同じ主題に属するコードの正規表現は、「または」文（'|'）で 1 つの正規表現に結合）。主題の正規表現が 1 つ以上マッチしていれば、その主題は対象文書に存在していると判断された。

さらに、欧州評議会加盟国 47 か国における倫理原則や価値観と、オブザーバー国やその他の国々における倫理原則や価値観の違いを比較している。

結論としては、世界各地で公表されている様々な AI に関するソフトロー（原則、ガイドライン等）には重要とされる価値について見かけ上の合意しかないとの指摘がある。また、項目の集計に加え、各項目の論点が精査されている。

2.3　総務省の報告書 2022 概要

この調査は、総務省の 2 つの AI 原則（AI 開発ガイドライン及び AI 利活用

ガイドライン）の見直しに関する論点を整理するために行われ、近年の国内外の動向や議論を踏まえた対応のため、これら 2 原則と国内外の他の原則の比較が行われた。16 か国 67 の原則（政府機関 40、業界団体等 10、グローバル企業 17）及び 22 の国内企業と経済団体が調査され、22 項目の星取表が作成された。見直しに関する論点として挙げられたのは、基本理念における「多様性」、「持続可能性」の追加の是非、「堅牢性」、「責任」、「追跡可能性」及び「モニタリング、監査」の項目としての追加の是非並びにパンデミック・災害等の有事を想定した見直しの是非である。

3　本研究の分析手法

CAHAI 報告書と報告書 2022 を比べると、刊行時期でみれば報告書 2022 が新しいが、CAHAI 報告書が学術論文の形式に則り分析の手法や結果について詳細な説明を設けているのに対し、報告書 2022 は政策検討を第一の目的とする資料調査の性質が強い。報告書 2022 では、総務省の 2 原則の項目と同じ内容を含む他機関の原則の例を数件ずつ挙げ、当該他機関の原則にも同じ論点が示されていることを指摘している。しかし、あくまで例示に留まり、当該論点が調査対象となった様々な原則の中でどの程度普遍的であったのかを読み取ることが難しいため、日本が国際的な潮流と足並みを揃えているのか、不十分な点があるかを判断する根拠として十分とは言いがたい。

報告書 2022 を批判的に論じたが、今後日本における AI 原則に関する検討を視野に入れた場合、総務省の既存の 2 原則については過去の政策に説明責任を有する行政の立場としては見直しになるであろうから、まさに報告書 2022 が射程としたように、当該既存 2 原則について分析をする必要がある。そこで本研究では、学術的な分析と政策面の需要を両立させるため、CAHAI 報告書の分析結果を用いて、総務省の 2 原則と国外諸原則との比較分析を実施する。

具体的には、次の段階により主題分析を実施した。

① CAHAI 報告書が示す 11 項目（4.1 にて後述）ごとに、関連する用語を特定
② ①で得た用語を日本の政府機関または日本企業の英訳版原則に採用されている訳語を用いて和訳

③　総務省 2 原則から②で得た和訳用語が出現する部分を抽出
④　②の和訳用語が出現しない部分についても 2 原則のテキストと②の和訳用語を比較して確認し、同義と考えられた部分を追加で抽出
例：「恵沢がすべての人によってあまねく享受」は「包摂」と同義
⑤　③④で抽出した各テキストを CAHAI が示す 11 項目に分類
⑥　分類した結果を CAHAI の分析結果と比較し、各テキストが 11 項目の内容に該当することを確認

加えて、総務省が公表する各国の AI 原則の調査を含む、「報告書 2022」第 3 章（総務省 , 2022）に示される AI 開発ガイドライン及び AI 利活用ガイドラインに関するレビュー及び見直しに関する論点整理の調査結果、国際機関が公表した現時点で最新の原則であるユネスコが 2021 年に公表した AI 倫理勧告、及び、国際機関として初の AI ガイドラインを公表し、各国の取組の情報共有を進めるための「AI 政策に関するオブザーバトリー」（通称「OECD.AI」）の取り組みを行う OECD が公表した報告書（OECD, 2022）にも触れながら、分析を行う。

4　分析内容（CAHAI 報告書と総務省 2 原則の比較）

4.1　各項目の分析

CAHAI 報告書で用いた Jobin, Ienca & Vayena の調査（Jobin et al., 2019）が特定した 11 の項目：透明性、正義と公正、危害の防止、責任、プライバシー、受益、自由と自律、信頼、持続可能性、尊厳、及び連帯とその論点が、総務省の 2 原則に項目として記載があるか、また論点との差異があるかを比較した。総務省の原則は、この 11 のうち 10 について言及していた。言及のない連帯（solidarity）は、EU の地域特性の強い項目である。また、総務省 2 原則にはこれら 11 の項目に対応しない「連携の原則」と「国際協力」が存在する。表 1 に、11 の項目と総務省独自の 2 項目について、論点の対応状況を示す。

表 1　CAHAI の論点と開発ガイドライン・利活用ガイドラインの論点対応

	CAHAI が抽出した論点	開発ガイドライン	利活用ガイドライン
透明性	(1) アルゴリズムとデータ処理方法の透明性 (2) 開発、展開に関連する人間の実践の透明性	②透明性の原則　(1) 入出力の検証可能性及び判断結果の説明可能性	⑨透明性の原則　(1) ①適正利用の原則　(2) 説明可能性を有する AI による人間の判断の実効性の確保
公平性	(1) AI 設計時及び導入時に、多様性を尊重し、インクルージョンと平等を支持 (2) 上訴や異議申し立てを行う可能性を求め、救済する権利 (3) AI の恩恵を受けるための公正な閲覧 (4) AI が労働市場に与える影響や、民主主義や社会的な課題に対処	基本理念　(3)「恵沢がすべての人によってあまねく享受」 ⑦倫理の原則　(3)「AI システムの学習データに含まれる偏見などに起因して不当な差別」が生じないように ⑧利用者支援　(1) 利用者に選択の機会を提供	基本理念　(3) 多様性を尊重、多様な背景と価値観、考え方を持つ人々を包摂 ①適正利用の原則　(1) 公平な条件による利用 ⑧公平性の原則　(1) サービスの判断のバイアス ②適正学習の原則　(1) データの質、アルゴリズムのバイアス
危害の防止	(1) 一般的な安心・安全 (2) サイバー戦争や悪意あるハッキングによる意図的誤用 (3) 社会的差別、プライバシー侵害、身体的・心理的危害 (4) 軍事目的での複製（デュアルユース） (5) 危害の軽減―AI の研究、技術的解決策、強制的なガバナンス介入 (6) データ品質評価やセキュリティ、プライバシー・バイ・デザイン、業界標準	④安全の原則　(1) (3) ・検証及び妥当性の確認 ・本質安全や機能安全 ・設計の趣旨及びその理由説明 ⑨プライバシーの原則 (2) (5) (6) ・情報の機密性、完全性及び可用性の確保、AI システムの信頼性や頑健性に留意。 ・検証・妥当性確認 ・セキュリティ・バイ・デザイン	①適正利用の原則 (2) (4) (5) ・利用者の信頼性の事前確認 ・誤用、悪用 ②適正学習の原則 (6) ・学習データの質の確保 ④安全の原則　(1) (3) ・生命・身体・財産に危害を及ぼさない被害想定、消費者への情報提供 ⑤セキュリティの原則 (5) (6)
責任・説明責任	・最終的な責任を負うのは常に人間だけであるべきか	情報提供・説明・フィードバック	サービスの提供者がステークホルダに対し情報提供、説明、対話

プライバシー	(1) 差分プライバシー、安全なマルチパーティ計算、同型暗号化などの技術的解決策 (2) 公共関与の解決策 (3) 法令遵守要件の明確化、新法や規則を作る等の規制アプローチの解決策	⑥プライバシーの原則 (1) (3) ・国際的な指針の踏襲 ・プライバシー影響評価 ・プライバシーデザイン	①適正利用の原則 (2) ・予防措置と事後対応 ・情報提供 ⑥プライバシーの原則 (1) (3) ・一般的配慮 ・学習モデル作成時のデータ提供・秘匿性の高い情報をむやみに AI に与えない
受益	受益の対象 民間企業：顧客の AI の利点 学界や政府機関：「すべての人」「人類」「社会全体」 (1) 人間の価値観に適合 (2) 権力の集中の抑制 (3) 人権促進	(1) 利用者の利益保護、リスク波及の抑止、人間中心の智連社会の実現	基本理念　(1) (2) (3) 民主主義社会の価値尊重、権利利益の侵害リスク抑制、便益とリスクのバランス確保
自由・自律	(1) 表現の自由 (2) 情報的自己決定 (3) プライバシーの権利 (4) 個人の自律性 (5) 実験、操作、監視からの自由	基本理念 (1) (4) 1 人間の尊厳と個人の自律 3 イノベーティブでオープンな研究開発と公正な競争と学問の自由や表現の自由	⑥プライバシーの原則 (2) (3) 本人同意なくパーソナルデータが第三者に提供されない ⑦尊厳・自律の原則 (4) (5)
信頼性	(1) 科学者や技術者の信頼文化の促進 (2) AI に対する過度の信頼への警告		①適正利用の原則 利用者の信頼性 最終利用者が誤って、悪意をもって使用していないか ⑩アカウンタビリティの原則 (1) プロバイダ及びビジネス利用者：情報提供と説明・ステークホルダとの対話
持続可能性	エネルギー効率の向上を要求 AI による雇用の喪失	「様々な課題の解決に大きく貢献」	「様々な課題の解決に大きく貢献」

尊厳	(1) 尊厳は人間の特権 (2) 人権の保護と推進 (3) AI は人間の尊厳を尊重、維持、向上	基本理念 1 (1) (3) 人間の尊厳と個人の自律が尊重される人間中心の社会 ⑦倫理の原則 (2) 「開発者は、人間の尊厳と個人の自律を尊重する。」	基本理念 (1) (3) 人間の尊厳と個人の自律が尊重される人間中心の社会 ⑦尊厳・自律の原則 (2) (3) ・意思決定・感情の操作等への留意 ・AI と人間の脳・身体を連携する際の生命倫理等の議論 ・プロファイリングを行う場合の不利益への配慮
連帯	(1) 労働市場にもたらす影響 (2) 弱者を尊重	—	—
国際協力	—		
AI 間の連携	—	①連携の原則 - ・情報の共有 ・国際的な標準や規格の準拠 ・データ形式の標準化、インターフェースやプロトコルのオープン化 ・相互接続・相互運用により意図しない事象が生ずるリスク ・知的財産権のライセンス契約及びその条件についてオープンかつ公平な取扱い	③連携の原則 AI がネットワーク化することによってリスクが惹起・増幅される可能性

表 1 の各項目については、以下 4.1.1 から 4.1.13 までのとおり分析した。

4.1.1　透明性

CAHAI 報告書によれば、透明性は 116 の文書のうち 101 で取り上げられ、他の原則のための倫理的条件と位置付けられているため出現数が多い。言及は (1) アルゴリズムとデータ処理方法の透明性と (2) AI システムの設計、開発、展開に関連する人間の実践の透明性に大別される。この 2 点は、日本の 2 原則にも見られる。また、透明性は「他の原則のための倫理的条件」とする記

述が利活用ガイドラインにあり、上記 2 つの論点を網羅しているといえる。

4.1.2　公平性

CAHAI の分析では、公平性、包摂性、差別につながるアルゴリズムバイアスの防止、民主主義や社会的な課題に対処する必要性が論点として抽出されている。技術的な部分に起因した差別の防止はどの地域でも重要とされているが、総務省の 2 原則では、すべての人が AI の恩恵を受けられるように、という包摂性が強調される。また、多様性について、「AI をめぐる議論の多様性を確保しつつ…認識の共有を図り」として、AI に関する議論自体の多様性を 2017 年の時点で強調していた点は特筆すべきである。この点は CAHAI 報告書においても「文化的多様性と道徳的多元主義の尊重と国家間の調和の必要性を両立させることが必要」(CAHAI, 2020, p.20) と重要性が指摘されている論点である。

一方、日本の原則では、一般的には情報格差等で弱い立場に置かれるとされる消費者 [3] に対し、異議申立て等の権利付与に触れることなく、主体的な行動を採るよう求める文言であり、今後見直しが行われるとすれば、1 つの論点となり得る。

4.1.3　危害の防止

本項目は、AI は予見可能な危害を起こしてはならないという意味で用いられる。CAHAI 報告書では、デュアルユースに言及しつつ、ハッキング、サイバー戦争、AI 兵器など具体的な場面を想定しつつ技術的解決やガバナンスの強制介入などを求めているのに対し、日本の原則は、AI の信頼性及び利用者の信頼性に言及し、技術的な安全対策、情報提供による適正な利活用に焦点が当てられる。また、「①適正利用の原則」、「②適正学習の原則」といった、積極的な利活用を促す項目が採用されていることも日本の特徴である。加えて、日本の原則は開発者、利用者及び消費者の間で起こる事象に注意が向けられているが、ハッキング、サイバー戦争などの意図的な危害への言及は「物理的な攻撃や事故への耐性」(総務省 , 2017, p.10) への留意という表現に留まり CAHAI 報告書に比べ不明瞭である。さらに、報告書 2022 (総務省 , 2022)

でも指摘されているが、災害やパンデミックなどの緊急事態（同報告書はこれを「有事」とも呼称）に生じうる問題についての記載を欠いていることは、一つの検討課題であろう。

4.1.4　責任・説明責任

CAHAI の分析では、責任と説明責任という概念が定義されることはほとんどないと問題提起されている。日本の 2 原則はこの語を「アカウンタビリティ」というカタカナの表記で採用しており、利活用ガイドラインにのみ脚注で「判断の結果についてその判断により影響を受ける者の理解を得るため、責任者を明確にした上で、判断に関する正当な意味・理由の説明、必要に応じた賠償・補償等の措置がとれること」と説明する。日本の原則でも誰がそれぞれの段階において責任を負うべきかの指針はない。

また、responsibility の意味での責任については、AI は人間のように責任を負うべきか、最終的な責任を負えるのは常に人間だけか、という点で意見の相違が見られると CAHAI は報告している。日本の 2 原則も、AI の行動と決定に対して責任と説明責任を負う様々なステークホルダが挙げられているが、AI が責任を負えるかについては記載がない。AI の責任負担の可能性について一致した見解はなく、課題となっていると言えよう。

4.1.5　プライバシー

本項目では CAHAI 報告書、日本の 2 原則共に、GDPR や個人情報保護法の遵守など、AI 以前のデータの取扱いに関する記述が多い。CAHAI も「ほとんどの文書は一般的な用語でプライバシーに言及しており、AI の能力と新しいプライバシーの課題との間に明確な関連性を確立していない」（CAHAI, 2020, p.16）と指摘する。しかし、利活用ガイドラインには、「ペットロボットなどの AI に過度に感情移入すること等により、特に秘匿性の高い情報をむやみに AI に与えることのないよう留意」（総務省 , 2019b, p.21）という記載があり、AI への過度の愛着ゆえにプライバシーを自ら公開してしまうという AI 特有の危険について言及しており、特筆される。また、利用者の意識向上の教育という視点があり、自己責任の考え方が色濃く表れる。

4.1.6　受益

CAHAI 報告書では、民間企業は顧客にとっての AI の利点を強調する傾向があるのに対し、学界や政府の関係者においては通常、AI は「すべての人」「人類」「社会全体」に利益をもたらすべきと主張されるという。その上で、同報告書では、主な論点として AI を人間の価値観に合わせること、権力集中の最小限化、人権促進のための AI 能力活用が挙げられている。総務省の 2 原則に受益は項目としては存在しないが、開発ガイドラインでは基本理念として「その恵沢がすべての人によってあまねく享受され、人間の尊厳と個人の自律が尊重される人間中心の社会を実現すること。」とあり、社会全体が AI の恩恵を受けること、人間の利益のための AI であること、という論点で CAHAI 報告書との一致を見ている。

4.1.7　自由と自律

CAHAI 報告書は、この自由は AI に決定や行動を支配されない自由と説明し、表現の自由、情報的自己決定、プライバシーの権利、個人の自律性、実験、操作、監視からの自由といった点が論点と指摘する。解決策として、透明で説明可能な AI の追求、AI リテラシーの向上、インフォームドコンセントの確保、データ収集・拡散の積極的な自粛などを示す。CAHAI が挙げた論点は日本の 2 原則でも網羅され、人間の尊厳や個人の自律の尊重、AI による意思決定や感情操作、具体例としては脆弱性に付け込まれること、プロファイリングが挙げられている。プライバシーの項目で前述したペットロボットの過度依存の例は、脆弱性に付け込まれることの問題として自由と自律の項目ともかかわりが深い。日本に特徴的な点としては、「AI と人間の脳や身体を連携する際の生命倫理等の議論」(総務省 , 2019b, p.22) について、つまりサイボーグの倫理的問題についての配慮が示されている点がある。また、プライバシーの項目と同様、リスクに留意した利用、情報の確認など、自らサービス提供者への情報提供を求め、専門的な内容を理解して判断するという比較的高度な内容が望ましいこととして消費者的利用者に求められている。

4.1.8　信頼性

日本の原則はAI自体への信頼を社会として醸成するという方向性であるが、AIは「信頼」の対象になるか、AIに対する国民の信頼を醸成することが道徳的に望ましいかという根本的な問いの存在をCAHAI報告書は示す。また、「信頼性」という特性と「信頼」という概念とが頻繁に混同され、互換的に使用されている点、及び誰と誰の信頼関係かが特定されていないという問題が指摘されている。

4.1.9　持続可能性

日本の原則はAIによる社会問題の解決（総務省，2019, p.7）という抽象的な楽観論にとどまる。一方、CAHAI報告書にはエネルギー確保の問題など、AIが持続可能性と対立する場合があることを踏まえた記述がある。この点はユネスコのガイドライン（UNESCO, 2021, p.21）には含まれており、ウクライナ戦争に端を発するエネルギー確保の問題も後押しして現実的な議論が進んでいる。さらに、OECDが公表した報告書（OECD, 2022）も、間接的影響を含む環境影響に関するデータ収集強化、透明性及び公平性の改善などを通じて、AIがグローバルな持続可能性目標に適合する手段であることが政策上確保されねばならないと指摘する。

4.1.10　尊厳

CAHAI報告書では、尊厳への言及は人権の保護及び推進と強く結びつき、AIは人間の尊厳を尊重し、維持し、向上させるものであるべきと主張する。また、尊厳は人間の特権でロボットの特権ではないと解される。一方で日本の利活用ガイドラインでは、意思決定の尊重が重要視される。しかし、脚注において、「AIは人間の活動を支援するものである…AIを人間と同様に扱うべきではないと考える（すなわち、人間の尊厳と個人の自律を尊重する）」（総務省，2019b, p.21）という説明が加えられ、AIに対する人間の優越が尊厳の尊重と結びつけられている点はCAHAIの報告書と考えを同じくする。加えて、サイボーグのような人間の身体とAIとの連携や融合による身体拡張を倫理的にどう捉えるか、その場合の人間の尊厳とは、といった新しい分野の問題提

起を行った点が注目に値する。

4.1.11　連帯

連帯は欧州連合基本権憲章 (European Parliament, 2000) の 6 項目のうちのひとつであり、労働権、社会保障、環境保護、消費者保護、などの要素を含む。異なる背景を持つ欧州市民が一致団結するための理念ともいえる。CAHAI の報告書は、AI が人間の労働を脅かすことへの対応、AI の恩恵の再配分、弱者の尊重に言及する原則があると述べる。この用語は EU 独自の用法であり総務省の原則に同じ意味での「連帯」の項目はない。開発ガイドラインでは利用者支援の原則として、ユニバーサルデザインなどの社会的弱者の利用促進についての記述があるが、あくまで社会的弱者という概念に留まっており、連帯が包括する範囲とは異なると言えよう。

4.1.12　国際協力

総務省の開発ガイドラインには、ステークホルダ間における適切な役割分担の実現、指針やベストプラクティスの国際的な共有、国際的な議論を通じたガイドラインの見直し・改定が挙げられ、ステークホルダの役割が言及されている。しかし、国際協力を単独の項目とする原則は世界的な傾向としては稀である。欧州評議会 AI Initiatives が公表する 604 の原則が取り入れる 31 の項目をまとめた調査結果 (Council of Europe, 2022) にも、「国際協力」の項目は取り上げられていない。

4.1.13　AI 間の連携

総務省は AI に関する会議の名称にも「AI ネットワーク」という用語を採用する等 AI 間の連携に初期の段階から注目していたことが窺える。この項目では相互接続性と相互運用性を確保するための関連情報の共有、標準化、他の AI 等と接続・連携することで制御不能となる等、AI が連携することによりリスクが惹起・増幅される可能性への留意について論じられる。CAHAI の「連帯」の項目では AI が人と人との繋がりにもたらす影響について述べているのに対し、この「連携」は AI と AI の繋がりが引き起こす影響について論

じているという違いがある。

4.2　全体分析

本節では、前節の分析を踏まえ、項目横断的な全体を通じた傾向について、分析を行う。

CAHAI が分類した 11 項目に含まれる論点は、おおむね総務省の 2 原則でも網羅されており、大きく欠けたものはなかった。特に透明性、プライバシー、受益に関しては、論点をほぼ同じくしている。しかし他の項目にいくつか差異が認められ、それらは、①地域の状況による差異、②考え方の多様性による差異、及び③時代の変化や議論の成熟の結果として出現した差異、の 3 種類に分類できる。具体的には以下のとおりである。

①地域の状況による差異

この点は、高い失業率や移民の問題を抱える地域で見られる「AI に仕事を奪われる」という懸念に関連した論点が日本では見られない点や、軍事転用のデュアルユースの問題について取り上げられていない点が挙げられる。

②考え方の多様性による差異

今後倫理の多様性として議論を深める余地があるものとして、AI を単なるモノととらえるか、それ以上になる可能性を秘めたものとして考えるかの違いから生じる論点がある。つまり、AI が「責任」や「信頼」の主体になりえるかという点である。欧米に偏ったデータ収集となっている CAHAI の調査は、あくまでも AI はモノであり、責任の主体や信頼の対象とはなりえないという立場である。しかし、日本のガイドラインでは AI が主体や対象となる書き方がされている箇所がいくつか存在する。また、利活用ガイドラインには、プライバシーの項目において過度の感情移入による脆弱性に付け込まれる場面の想定がなされている点、尊厳の項目において、「AI を人間の脳・身体と連携させる場合、特に、エンハンスメント（健康の維持や回復を超えた人間の能力の増進の追求）を行う場合」（総務省 , 2019b）に人間の尊厳と自律が侵害されないよう留意することなど、より高度な技術が身近になる近未来を想定し

た記載があることが特筆される。

また、2017 年の時点で「開発者、市民社会を含む利用者など関係するステークホルダは…AI をめぐる議論の多様性を確保しつつ…認識の共有を図り、相互に協力するよう努める」(総務省 , 2017) として、AI の議論の多様性の確保を掲げている開発ガイドラインの内容は世界に先駆けた取り組みであったと言えよう。

③時代の変化・議論の成熟の結果として出現した差異

総務省の利活用ガイドラインは 2019 年、CAHAI 報告書が確認した文書は 2020 年初頭までの公表のものであるため、COVID-19 を念頭に置いた記載や、米国の非営利団体「OpenAI」により 2022 年 11 月に公表された ChatGPT をはじめとする生成 AI に関する具体的な課題は想定されていない。パンデミック等有事の対応については既に報告書 2022 において見直しの論点として示されており、パンデミックのほか、大規模災害や戦時下などの有事と AI の利用に関する記載という観点から、検討を要すると考えられる。

また、「持続可能性」の扱いについても、総務省の 2 原則では AI が国際的な問題の解決に寄与する、といったプラス面での記載のみとなっているが、近年はマイナス面もあることが認識されている。主に莫大なエネルギーの確保や差別、格差の助長といった面で AI の発展が持続可能性と一部対立する点について CAHAI の報告書も指摘している。2021 年に公表されたユネスコの勧告においても、「AI 技術の出現は、開発レベルの異なる国々でどのように適用されるかによって、サステナビリティの目標に利益をもたらすこともあれば、その実現を阻害することもあり得る」(UNESCO, 2021, p.21) として、AI が持続可能性と対立する可能性を指摘している。加えて、2022 年公表の OECD 報告書 (OECD, 2022) は、より具体的に、AI の環境影響について計測手法の改善が必要と指摘しつつも、効率性を高める汎用技術としての側面における間接的な環境影響はプラスになり得るが、資源を利用する側面における直接的な環境影響はほぼマイナスであると評価している。

表２　CAHAI 報告と総務省の２原則の論点の差異

	論点一致	CAHAI のみの論点	日本独自の論点	差異がある／議論が必要な論点
透明性	◎			
公平性	○	AI の労働市場への影響	AI の議論の多様性	
危害の防止	△	・サイバー戦争等の意図的な危害 ・軍事目的の利用		緊急事態（有事）の対応
責任・説明責任	◎			責任主体は誰か、AI は責任主体となり得るか
プライバシー	◎		AI への過度な信頼により自らプライバシーを明かしてしまう危険	AI 特有の問題についての議論
受益	◎			
自由・自律	◎		ペットロボットなどへの過度な感情移入	
信頼性	△			・「信頼性」と「信頼」の違いの認識 ・AI は「信頼」の対象となり得るか
持続可能性	△			AI が持続可能性と対立する場合の認識とバランス
尊厳	○		AI と人間の脳・身体を連携する差異の生命倫理	
連帯	―			
国際協力	―			
AI 間の連携	―		AI の連携によるリスクの増大	

5　結論及び今後の政策検討に向けた示唆

本研究では、欧州評議会における分析結果や国際機関の情報源を用いて総務省の２つの AI 原則に対して分析を行い、主に次の知見を得た。

まず、総務省の２原則は、AI 原則としては初期に策定されたものでありながら今なお世界の AI 原則に関する主要な論点をおおむね網羅していることが

確認された。特に、AI によって少数派が不利益を被らない、という意味の多様性のみならず、AI の社会受容や AI 倫理の議論における多様性の確保を当初から掲げていた点が特筆される。また、AI をあくまでもモノとして扱う欧米の諸原則[1)]と異なり、AI が主体や対象となる記載があること、AI への過度な感情移入及び AI と脳・身体の間の連携など、近未来的状況を想定した記載があることも特徴である。これらの点は、多様性の確保を目指す国際社会に対して、今後さらに訴求していくことが可能な点と考える。

一方で、総務省 2 原則には、緊急事態対応を含む「危害の防止」や「持続可能性」などで追加を検討するべき論点があることも明らかになった。これらの点について、本研究の分析で得た内容に考察を加えつつ述べると、次のとおりである。

まず緊急事態であるが、自律型兵器など、軍事目的での AI 利用を推進している国々と日本の議論は対局にある。安全保障の観点からは国内においてほぼ議論されておらず、防災や災害発災時における緊急事態への対応及び感染症対策との関係で、若干の議論があるに過ぎない。安全保障の観点から緊急事態を捉え検討を進めている諸外国とは議論と整合性という観点で、同じ土俵で議論をしている状況にはない。国民の生命、身体及び財産の保護等のために必要不可欠な要素（内閣官房 , 2020）である安全保障を直視した検討が行われることを強く期待したい。

プライバシーについて、EU において準拠すべき法令である GDPR への準拠は当然の帰結といえよう。一方で日本においては、個人情報保護法がプライバシーの権利を直接に明記し保障していないため、AI に関連する個人情報の保護の問題は EU におけるプライバシーの権利を基軸とする取り組みとは一線を画す。さらに、日本は、AI へ過度の信頼によって、自らのプライバシーを晒してしまう危険など、自らの情報をコントロールすることについて、独自の懐疑的な視点もある。プライバシーは、国境を越えた情報の流通を左右する要素であるとともに、各国で異なる個人の意識と関係する部分も大きく（総務省（2020）第一部第 3 章第 3 節；大磯ら（2021））、国際的な動向を踏まえながら、日本に適合した解決法を採っていくことが妥当であろう。

持続可能性について、日本は AI で課題を解決できるという方向での議論が

なされてきた。一方で、本論文の分析の通り、社会においてAIを実装することが持続可能性の阻害要因になり得るという指摘が、近年ユネスコやOECDからなされている。国際機関並びに各国政府は、新興技術の開発に伴って生ずる新たな問題や環境への対応も含めて、今後さらに研究や検討を進める必要があると考えられる。

以上、総務省2原則に対して追加を検討すべきと考えられる論点を示した。現実の政策検討としては、原則の具体的な見直しについて実施されるという発表はまだされていない[4)]。本研究結果が今後の何らかの検討の材料となれば幸いである。また、学術研究の視点としても、特に、持続可能性、生成AI特有の問題など本稿において検証し議論を尽くすことができなかった課題について今後の更なる研究の発展が期待できることを指摘して、結びとしたい。

謝辞

本研究は、JSTムーンショット型研究開発事業、JPMJMS2215の支援を受けたものです。

注

1) CAHAIの報告書では、「尊厳」の項目において尊厳は人間の特権であり、ロボットの特権ではないと明言している。知的活動体としての「尊厳」は認めておらず、あくまで機械という扱いである、といった点。

2) 2022年2月8日に開催の総務省AIネットワーク社会推進会議（第20回）AIガバナンス検討会（第16回）合同会議において、「AI開発ガイドライン及びAI利活用ガイドライン に関するレビュー（1）」が、続く4月27日のAIネットワーク社会推進会議（第21回）AIガバナンス検討会（第17回）合同会議において、AI開発ガイドライン及びAI利活用ガイドラインに関するレビュー（2）」及び、「報告書2022（骨子）」（案）が公表された。メール等による審議を経て、「報告書2022」が2022年7月25日に公表された。

3) OECDのConsumer Policy Toolkitによれば、'The market failures that arise out of the lack of information are a primary focus of consumer protection policy'. とあり、情報が不足している消費者は保護されるべき対象と見なされている。OECD（2010）Consumer Policy Toolkit, p.32, OECD Publishing, Paris,〈https://doi.org/10.1787/9789264079663-en〉

4) 本記載は本稿執筆中の2022年11月現在の状況である。2023年5月20日、G7広島サミットにて、「G7広島首脳コミュニケ」（https://www.mofa.go.jp/mofaj/files/100507033.pdf）が発出され、広島AIプロセスを年内に創設するとし、10月30日に「広島AIプロセスに関するG7首脳声明」（https://www.mofa.go.jp/mofaj/files/100573465.pdf）を発出した。また、に2023年9月8日に内閣府AI戦略会議

より「新 AI 事業者ガイドライン　スケルトン（案）」（https://www8.cao.go.jp/cstp/ai/ai_senryaku/5kai/gaidorain.pdf）が公表された。

参考文献

大磯一、依田高典、黒田敏史（2021）「個人のプライバシー意識等とデジタルサービス利用に関する実証分析」『情報通信学会誌』39（3）, pp.37-47. https://doi.org/10.11430/jsicr.39.3_15

カーツワイル , レイ（2007）『ポスト・ヒューマン誕生ーコンピュータが人類の知性を超えるとき』NHK 出版 .

新保史生（2020）「AI 原則は機能するか？―非拘束的原則から普遍的原則への道筋」『情報通信政策研究』3（2）, pp.53-70.

総務省 AI ネットワーク社会推進会議（2017）「国際的な議論のための AI 開発ガイドライン案」https://www.soumu.go.jp/main_content/000499625.pdf（2022 年 11 月 22 日アクセス）

総務省 AI ネットワーク社会推進会議（2019a）「報告書 2019 概要」https://www.soumu.go.jp/main_content/000637103.pdf（2022 年 11 月 30 日アクセス）

総務省 AI ネットワーク社会推進会議（2019b）「AI 利活用ガイドライン ～ AI 利活用のためのプラクティカルリファレンス～」https://www.soumu.go.jp/main_content/000637097.pdf（2022 年 11 月 21 日アクセス）

総務省 AI ネットワーク社会推進会議（2022）「報告書 2022」https://www.soumu.go.jp/main_content/000826564.pdf（2022 年 11 月 21 日アクセス）

総務省（2020）「情報通信白書令和 2 年版」https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/html/nd133410.html（2022 年 11 月 28 日アクセス）

内閣官房国土利用の実態把握等に関する有識者会議（2020）「提言」https://www.cas.go.jp/jp/seisaku/kokudoriyou_jittai/index.html（2023 年 6 月 14 日アクセス）p.9

内閣府統合イノベーション戦略推進会議（2019）「人間中心の AI 社会原則」https://www8.cao.go.jp/cstp/aigensoku.pdf（2022 年 11 月 21 日アクセス）

中川裕志（2020）「AI 倫理指針の動向とパーソナル AI エージェント」『情報通信政策研究』3（2）, pp.1-23.

福岡真之介（2020）「AI の責任と倫理（第 1 回）AI 倫理原則の世界的動向」『NBL』1168, pp.49-58.

Arksey, H., O'Malley, L. (2005) "Scoping studies: towards a methodological framework", *International journal of social research methodology*. 8(1), pp.19-32.

CAHAI (2020) AI Ethics Guidelines: European and Global Perspectives, Ad Hoc Committee on Artificial Intelligence, Council of Europe, https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac（2022 年 11 月 30 日アクセス）

Council of Europe (2022) Datavisualisation of AI Initiatives, https://www.coe.int/en/web/artificial-intelligence/national-initiatives（2023 年 3 月 31 日アクセス）

European Parliament (2000) The charter of Fundamental Rights of the European Union, https://www.europarl.europa.eu/charter/pdf/text_en.pdf（2023 年 3 月 28 日アクセス）

Hagendorff, Thilo (2020) "The Ethics of AI Ethics:An Evaluation of Guidelines", *Mind and Machines.* 30, pp.99-120.

Jobin, A., Ienca, M., Vayena, E. I. (2019) 'Cornell University Library, Artificial Intelligence: the global landscape of ethics guidelines' *Cornell University Library, arXiv.org*

OECD (2019) Recommendation of the Council on Artificial Intelligence 人工知能に関する理

事会勧告 , https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449（2023年4月3日アクセス）
OECD (2022) Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint, OECD Digital Economy Papers, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en
UNESCO (2021) Recommendation on the ethics of artificial intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000381137（2022年11月24日アクセス）

〔受付日　2022. 11. 30〕
〔採録日　2023.　7.　3〕