2.1.9 社会におけるAI

（1）研究開発領域の定義

人工知能（AI）技術が社会に実装されていったときに起こり得る、社会・人間への影響や倫理的・法的・社会的課題（Ethical, Legal and Social Issues：ELSI）を見通し、あるべき姿や解決策の要件・目標を検討し、それを実現する制度設計および技術開発を行うための研究開発領域である。

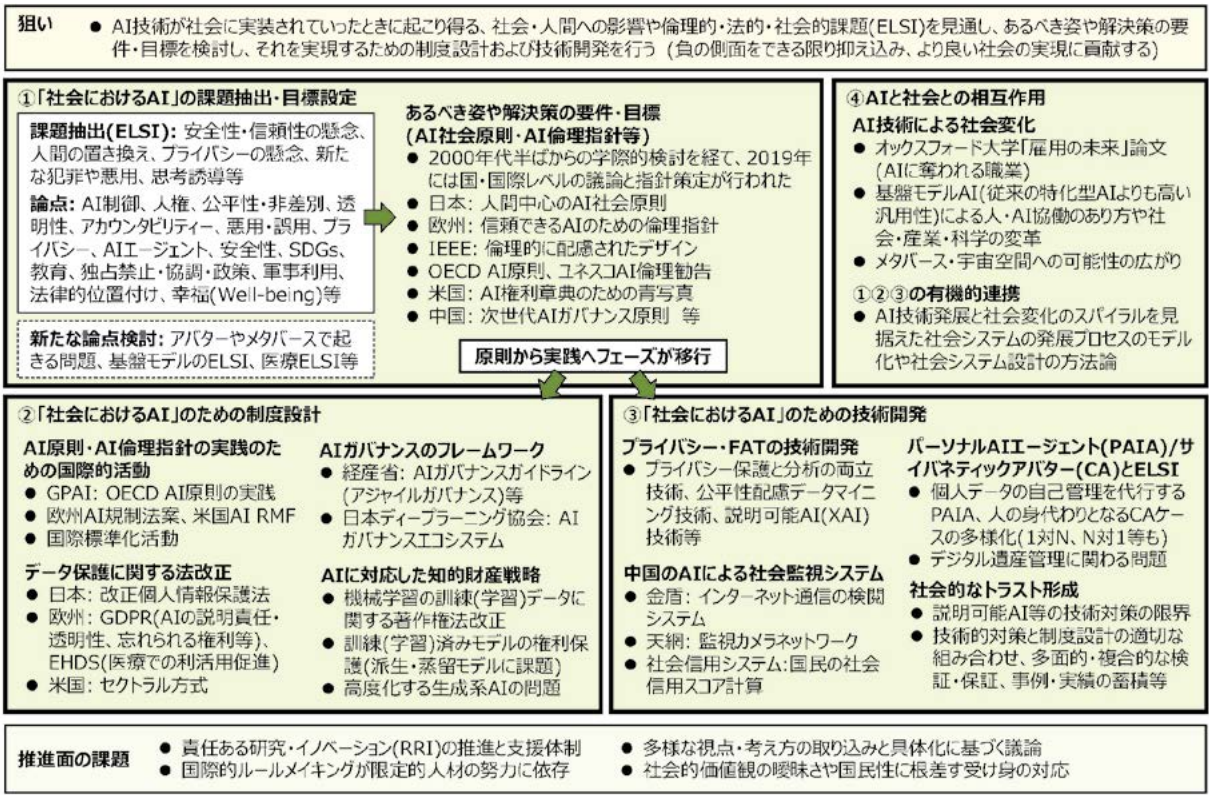

図2–1–13　　領域俯瞰：社会におけるAI

（2）キーワード

ELSI、RRI、FAT、AI倫理、AI社会原則、公平性、アカウンタビリティー、透明性、トラスト、ガバナンス、法制度、プライバシー、知的財産権、AIエージェント

（3）研究開発領域の概要

[本領域の意義]

AI技術は、人間の知的作業をコンピューターで代行する可能性を広げることで、社会における人間の役割を変え、人間の働き方やモチベーションにも影響を及ぼし、社会の仕組み・在り方も変貌させる可能性を持っている。これによって、便利で効率的な社会を築くことができ、人間は快適な生活を過ごせると期待される一方で、AIが職業を奪うとか、プロファイリング（個人の性格・特徴を分析する技術）によってプライバシーを侵害されるとか、負の側面に対するさまざまな不安・懸念が指摘されている。本研究領域の取り組みは、それら起こり得る影響・課題を事前に把握し、その対策を制度と技術の両面から実現することによって、負の側面をできる限り抑え込み、より良い社会を実現するために貢献する。

［研究開発の動向］

本領域の取り組みを、❶「社会におけるAI」の課題抽出・目標設定、❷「社会におけるAI」のための制度設計、❸「社会におけるAI」のための技術開発、❹AIと社会との相互作用、という四つに分け、その概要と動向を述べる。

❶「社会におけるAI」の課題抽出・目標設定

❶はAI技術が社会に実装されていったときに起こり得る、社会・人間への影響や倫理的・法的・社会的課題（ELSI）を抽出し、あるべき姿や解決策の要件・目標を定める活動である。

人間が行っていた知的判断のタスクがAIによって代替・自動化され、人間を上回る精度・規模・速度で処理されるようになってきた。これによって、さまざまなシーンで効率化・最適化、人間の負荷軽減がなされることは大きなベネフィットであるが、反面、AIによるタスク代替は人間の役割や心理に急激な変化をもたらすことでネガティブインパクトも生む。これがAIのELSIとして論じられている[1), 2), 3), 4)]。その代表的なものとして、以下のような問題が挙げられる。

- **安全性・信頼性の懸念：**機械学習は原理的に動作保証や精度保証が難しいこと（品質保証問題）、結果についての理由説明がされないこと（ブラックボックス問題）、偏見・差別を含んだ学習をしてしまうこと（バイアス問題）、Adversarial ExamplesのようなAI特有の脆弱性が存在すること（脆弱性問題）など、システムの安全性・信頼性に対する懸念が指摘されている[5)]。
- **人間の置き換え：**従来は人間が行っていたタスクがAIによって自動化されることで人間の失業が増えるという懸念、大量に生まれ得るAIによる生成物に関わる著作権の問題、AIによって故人（のある一面）を複製する行為の倫理問題、擬人化されたAIエージェントに心理的に依存するケースなどの懸念が指摘されている。
- **プライバシーの懸念：**さまざまな行動履歴データを解析することで、個人行動が追跡されやすい状況であることや、映像解析やバイオメトリクス解析によって個人の感情・心理状態などが読み取れるようになりつつあることなど、AI技術の発展に伴うプライバシー侵害の懸念が高まっている。
- **新たな犯罪や悪用：**AI技術を用いることで、本物と区別困難なフェイク画像・音声・映像が簡単に生成できるようになり、まるで人間が書いたかのような自然な文章生成や対話応答が可能になったことで、なりすましや偽装への悪用や、詐欺のような犯罪行為の巧妙化を招いている。
- **思考誘導：**AIによるリコメンデーションへの依存が高まると意思決定の主体性が低下していく懸念、情報のパーソナライズやソーシャルネットワークにおけるフィルターバブルやエコーチェンバー現象、フェイクニュースを用いた政治操作・プロパガンダなど、人々の思考が誘導されやすいというリスクが高まっている[6), 7)]。

次に、このような問題を議論し、あるべき姿や解決策の要件・目標を定めようとする取り組みが国内外で推進されている[1), 8), 9), 10), 11), 12), 13)]。以下にその代表的なものを挙げる。また、これらで重視されている論点を表2–1–3に挙げた[1]。

1 表2–1–3の論点項目は、主要なAI倫理指針を参照して文献7）で整理されたものである。同文献では、主要なAI倫理指針のそれぞれでどの項目が重視されているかについても比較表にまとめている。

表2–1–3　AI ELSIの主要な論点[11)]

論点	説明
AI制御	AIは人間によって制御可能でなくてはならない
人権	AIは人権を尊重するように設計されるべき
公平性・非差別	AIの処理結果によって、人々が不当に差別されないように配慮すべき（主にAIが用いるデータやアルゴリズムにバイアスが含まれることに起因する）
透明性	AIの動作の仕組みは開示されるべき、AIの動作の仕組みや処理結果は人々が理解できるレベルで説明可能であるべき
アカウンタビリティー[2]	AIが事故などを引き起こした際に、その原因や責任の所在を明らかにできるべき
トラスト	AIは人々が信頼できるものであるべき（AIの動作を予想できるとか、処理結果を受容できるとかいったことを含む）
悪用・誤用	AIの悪用・誤用を防ぐような対策を考えるべき
プライバシー	AIの開発時・利用時に人々のプライバシーを侵害してはいけない（開発時の学習データの個人属性や、利用時の内面や機微な情報に立ち入る分析など）
AIエージェント	AIエージェントは、そのユーザーの個人データの管理代行をするが、ユーザーの意思に沿った処理（プライバシー保護も含む）を行わねばならない
安全性	AIはそのユーザーおよび他の人々の生命・身体・財産などに危害を及ぼさないように設計されるべき
SDGs	SDGsで掲げられているような環境・社会などの課題にAIによる貢献を目指す
教育	AIについての理解や倫理・リテラシーを含む分野横断・学際的教育が求められる
独占禁止・協調・政策	特定の企業や国によるAI技術やデータ資源の独占は望ましくない、人材・研究の多様化・国際化や産学連携、国際協調・開発組織間協調が望まれる
軍事利用	自律型致死兵器システム（LAWS）に代表されるAIの軍事利用を制限すべき
法律的位置付け	AIを法律的にどのように位置付けるべきか（例えばAIに人格権などを与えるか）
幸福（Well-being）	AIは人々の幸福のために用いる

欧州では、比較的早い時期から、特に英国の大学・研究機関を中心に取り組まれてきた。まず、2005年にオックスフォード大学の哲学科の下部組織としてFuture of Humanity Institute（FHI）が設立された。FHIは、技術変化によってもたらされる倫理的ジレンマやリスクに対して、長期的にどう選択・対処していくべきか、学際的な研究を進めている。また、ケンブリッジ大学では、2012年に人文・社会科学部局の下部組織としてCambridge Center for Existential Risk（CSER）が設立された。CSERでは、AI、バイオ、ナノなどの先端技術のリスクに対する哲学的・倫理的な研究が行われており、産官学ワークショップなどを実施している。最近の重要な動きとして、2019年4月に欧州委員会のAI HLEG（High-Level Expert Group on Artificial Intelligence）が「信頼できるAIのための倫理指針（Ethics Guidelines for Trustworthy AI）」を公表した。さらに、欧州委員会は2020年2月に「AI白書」（White Paper on Artificial Intelligence - A European approach to excellence and trust）を発表して、市民の価値観と権利を尊重した安全なAI開発の「信頼性」と「優越性」を実現するための政策オプションを示し、2021年4月に「AI規制法案（Proposal for a Regulation of the European Parliament and of the

2　Accountability（アカウンタビリティー）の和訳として「説明責任」が用いられることが多い。「説明責任」という言葉から、説明すればよいと解釈されやすいが、Accountabilityには本来、説明に加えて、法的あるいは経済的な責任を取ることも含まれているということを踏まえておくべきである[4)]。

Council Laying Down Harmonised Rules on Artificial Intelligence（Artificial Intelligence Act）and Amending Certain Union Legislative Acts）」を公表するに至る。これは❷「社会におけるAI」のための制度設計の段階に入るものであり、その内容は❷にて後述する。

米国では、産業界や非営利組織が主導する形で取り組みが始まり、それを追うように学術界での取り組みや国の政策が立ち上がった。2014年3月設立のThe Future of Life Institute（FLI）、2015年12月設立のOpenAI、2016年9月設立のPartnership on AIなどの非営利組織がよく知られている。特にFLIは、2017年1月に5日間にわたるアシロマ[3]での会議の結果として、AIの研究課題、倫理と価値観、長期的な課題を含む23項目のガイドライン「アシロマAI原則」（Asilomar AI Principles）を公表し、多くの署名賛同を得ている[4]。2018年10月には、電子プライバシー情報センター（Electronic Privacy Information Center：EPIC）によって設立された団体であるPublic Voiceが「AIユニバーサルガイドライン（Universal Guideline for Artificial Intelligence）」を公表し、AIの設計や利活用の改善を目的として12の原則を提案した。 AIシステムに関わる主な責任は、同システムに資金を供給し、開発し、展開する機関にあるべきと言及している。

一方、米国の学術界での取り組みとしては、スタンフォード大学のOne Hundred Year Study on Artificial Intelligence（AI 100）、IEEE（The Institute of Electrical and Electronics Engineers：米国電気電子学会）の自律インテリジェントシステムの倫理に関するIEEEグローバルイニシアチブ（The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems）がよく知られている。特に注目されるのはIEEEグローバルイニシアチブでグローバルイニシアチブを開始し、「倫理的に配慮されたデザイン（Ethically Aligned Design）：自律インテリジェントシステムで人間の福祉を優先するためのビジョン」と題されたレポートを作成し、2016年12月にVer.1（EADv1）、2017年12月にVer.2（EADv2）を経て、2019年3月に1st Edition（EAD1e）をリリースした。 EADはデザインという言葉を使っている通り、倫理そのものではなく、設計論・設計思想、それをどのように技術に落とし込めるかといった論点が整理されていることが特徴である[14)]。 EADv2では自律型兵器システムのような問題にも踏み込んで論点を広げたが、最終的なEAD1eでは八つの原則に絞り込んだ。さらに、これらの原則を実践に結び付けるため、IEEE–SA（Standard Association：標準規格）のP7000シリーズとして標準化活動が進められている。米国政府からは2022年10月に「AI権利章典のための青写真（Blueprint for an AI Bill of Rights）」が公開された。

日本では、学会・政府主導のガイドライン策定が推進されている。学会では2014年に人工知能学会が倫理委員会を立ち上げ、同委員会での議論や公開討論を経て、2017年2月に「人工知能学会 倫理指針」を公開した。9項目から成り、主に研究者倫理に焦点が置かれているが、第9条「人工知能への倫理遵守の要請」はAI自体が倫理的であるべきということを掲げたのが特徴である。また、政府主導の活動としては、内閣府の「人工知能と人間社会に関する懇談会」、総務省の「AIネットワーク社会推進会議」[5]、経済産業省の「AI・データ契約ガイドライン検討会」などが進められてきたが、それらを踏まえた活動として、内閣府の「人間中心のAI社会原則検討会議」が2018年5月に始まり、2019年3月に「人間中心のAI社会原則」が決定・公表された。人間中心のAI社会原則は、人間の尊厳が尊重される社会（Dignity）、多

3 米国カリフォルニア州のアシロマは、遺伝子組み換えに関するガイドラインが議論されたアシロマ会議が、1975年に開催された場所である。このアシロマ会議は、科学者自らが研究の自由を束縛してまで自らの社会的責任を表明したもので、科学史に残る象徴的な場所で再びAIに関して同様の議論がなされた。

4 2023年2月2日時点の公開情報として、AI・ロボット工学研究者1797名、その他3923名がこの原則に署名したとのことである。

5 「人間中心のAI社会原則」に先立ち、2017年7月に「国際的な議論のためのAI開発ガイドライン案」、2018年8月に「AI利活用原則案」を公開している。後者はさらに「人間中心のAI社会原則」の発表後、2019年8月には「AI利活用ガイドライン」としてリリースされた。

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）という三つの価値を基本理念とし、「AI-Readyな社会」をビジョンに掲げ、人間中心の原則、教育・リテラシーの原則、プライバシー確保の原則、セキュリティー確保の原則、公正競争確保の原則、公平性・説明責任・透明性の原則、イノベーションの原則という七つをAI社会原則として挙げている。

AI原則に関して、2019年は国際的な協調が議論された年でもあり、経済協力開発機構（Organisation for Economic Co-operation and Development：OECD）は5月に「人工知能に関するOECD原則（OECD Principles on Artificial Intelligence）」をまとめ、42カ国[6]が署名した。6月に日本で開催されたG20貿易・デジタル経済大臣会合では、「人間中心」の考えを踏まえたAI原則「G20 AI原則」に合意がなされた。さらに、ユネスコ（国際連合教育科学文化機関、United Nations Educational, Scientific and Cultural Organization：UNESCO）での検討も2019年から始まり、2021年11月に「AI倫理勧告（first draft of the Recommendation on the Ethics of Artificial Intelligence）」が全193加盟国[7]によって採択された。この勧告では、AIを開発・利用する際に尊重すべき価値として「人権」「環境保全」「多様性」「平和や公正さ」を掲げ、プライバシー保護や透明性確保など守るべき10の原則を規定している。このような国際的な動き[15)]と連動するように、上に述べた以外にも各国からAI原則が発表された。中国では、2019年5月に北京智源人工智能研究院（Beijing Academy of Artificial Intelligence：BAAI）が「北京AI原則（Beijing AI Principles）」を公表、6月には中国国家次世代AIガバナンス専門委員会が「次世代AIガバナンス原則―責任あるAIの発展」を公表、さらに中国AI産業発展連盟が「AI業界自律公約」を定めた[8]。また、企業や企業グループが自社の取り組みとしてAI原則・AI倫理指針を掲げるという動きも国内外で広がった。

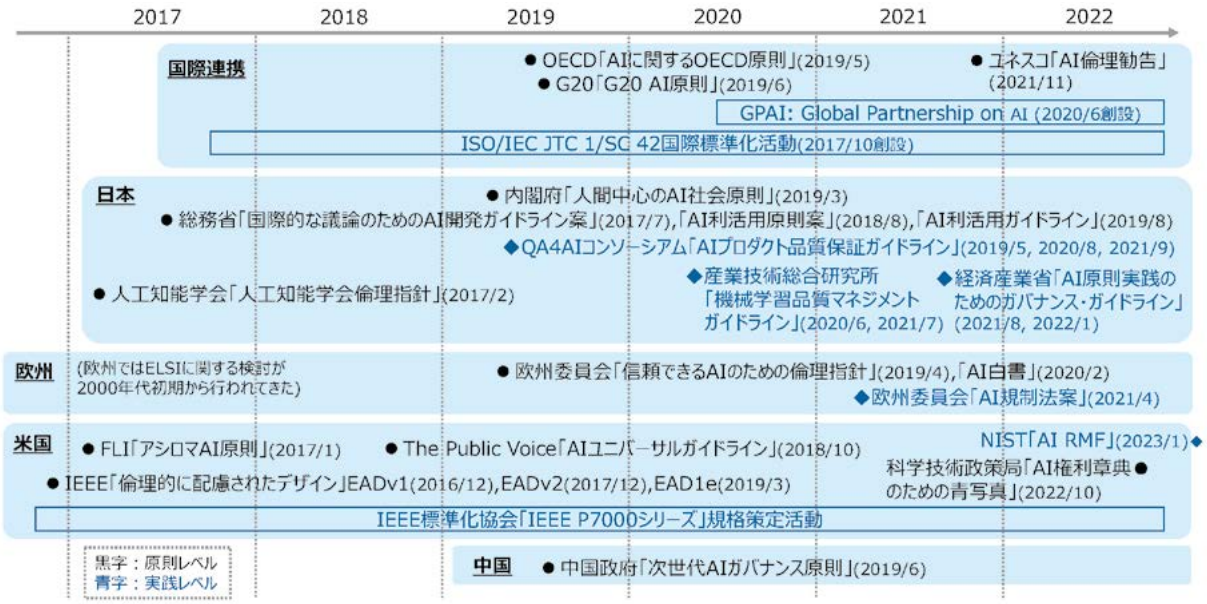

図2-1-14　AI ELSI関連ガイドラインを中心とした主要な取り組み[12)]

6　OECD加盟36カ国に、アルゼンチン、ブラジル、コロンビア、コスタリカ、ペルー、ルーマニアを加えた42カ国。

7　米国は含まれていない。中国は含まれている。

8　「次世代AIガバナンス原則」が国家戦略「次世代AI発展計画」を受けたもの、「北京AI原則」は北京の研究機関が中心となって発信したもの、「AI業界自律公約」は産業界の順守を期待するものとなっている。

以上に挙げたように、さまざまな国・組織からAI原則・AI倫理指針が出されたが、OECDやG20のような国際的な場での議論も行われており、それらで取り上げられている事項には共通点が多く見られる。現在は、このような理念・原則レベルの議論から実践のフェーズへと移行している。実践フェーズの取り組み内容は、このあと❷❸で紹介する。原則から実践へという全体の流れを、AI ELSI関連ガイドラインに関する取り組みを中心に図2–1–14に示した。

❷「社会におけるAI」のための制度設計

❶で導出した要件・目標の実現に向けて、制度設計面の取り組みが進められている。主な取り組みとして、a. AI原則・AI倫理指針の実践のための国際的活動、b. AIガバナンスのフレームワーク、c. プライバシー・個人情報保護などのデータ保護に関する法改正、d. AIに対応した知的財産戦略、が挙げられる。以下ではa～dそれぞれの動向について述べる。なお、ここでは取り上げないが、他に自動運転・自律飛行や医療AIといった個々のAI応用ごとの制度整備なども進められている[16)]。

❷ –a AI原則・AI倫理指針の実践のための国際的活動

❶に示したようなAI原則・AI倫理指針は国や組織でさまざまな形で実践に結び付ける取り組みが進みつつあるが、特に国際的な活動としてGPAI（Global Partnership on AI）が挙げられる。これは、前述の「人工知能に関するOECD原則」を実践段階に進めるための国際的な組織である。その詳細は［国内外の注目プロジェクト］❶で述べる。

また、欧州では前述の通り、欧州委員会が2019年4月に公表した「信頼できるAIのための倫理指針」から、2020年2月の「AI白書」を経て、2021年4月には「AI規制法案（AI Act）」を公表した。この法案では、AI応用システムをリスクの大きさに着目して四つのレベルに分け、そのレベルに応じて使用禁止や適合性評価の義務化など、かなり踏み込んだ規制をかけようとしている。さらに欧州では、人権・民主主義・法の支配を掲げる欧州評議会（Council of Europe：CoE）[9]のCAI（Committee on Artificial Intelligence）において、AI条約の起草が進められている。リスクベースの考え方に基づく枠組み条約を方針とし、2023年11月の採択を目指している。

一方、米国では、「2020年国家AIイニシアチブ法（National Artificial Intelligence Initiative Act of 2020）」を受けて、前述の「AI権利章典のための青写真」（2022年10月公開）に続き、「AIリスク管理フレームワーク（Artificial Intelligence Risk Management Framework：AI RMF）」が2023年1月に発表された。標準技術研究所（National Institute of Standards and Technology：NIST）から発表されたもので、AIのリスクに対する考え方やリスクに対処するための実務が示されている。

欧州のAI規制法案と米国のAI RMFについては、［新展開・技術トピックス］❶でもう少し詳しい内容を記載するが、これらは国家レベルの政策として、原則から実践へトップダウンに落とし込む流れである。それに対して日本では、産業界や研究開発の現場主体のボトムアップな取り組みによって、AIシステムの安全性・信頼性の確保のための方法論が検討され、具体的応用を踏まえた開発者目線の実践的なAI品質管理ガイドラインが作られている。 AIプロダクト品質保証（QA4AI）コンソーシアムによる「AIプロダクト品質保証ガイドライン」や、産業技術総合研究所による「機械学習品質マネジメントガイドライン」がその代表例である。これらについては「2.1.4 AIソフトウェア工学」の中で取り組みを紹介している。

上記のような欧州・米国の政策は、その国・地域内にとどまらず国際的に大きな影響力を持つ。ただし、これと並行して、AI倫理・AIガバナンスを含むAIに関する国際標準化活動が進められており、その中では、

9 欧州連合（EU）とは別の機構である。現在、46カ国が加盟しており（欧州で未加盟なのはベルラーシと2022年3月に除名されたロシアの2カ国のみ）、日本・米国など5カ国がオブザーバー国となっている。

欧州・米国だけでなく、中国・日本を含む各国の考え方を交えて活発な議論が行われている。標準化活動では、抽象的な理念だけでなく、システム開発に直結する面も大きいため、日本で検討してきた実践的なガイドラインも重要な貢献を示している。なお、標準化活動は開発方法論との関係が深いので「2.1.4 AIソフトウェア工学」に記載した。

❷ –b. AIガバナンスのフレームワーク

原則から実践へという動きは、国・国際レベルに限らず、個々の企業・組織の現場での実践が重要になる。これに関して、AIガバナンスという言葉がよく使われ、その実践のためのフレームワークやガイドラインの整備が求められている。

国内では、経済産業省から「我が国のAIガバナンスの在り方」（Ver. 1.1、2021年7月）、「AI原則実践のためのガバナンス・ガイドライン」（Ver. 1.1、2022年1月）、「AI・データの利用に関する契約ガイドライン」（Ver. 1.1、2019年12月）が公開されている。「我が国のAIガバナンスの在り方」では、AIガバナンスとは「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」と定義している。「AI原則実践のためのガバナンス・ガイドライン」は、企業ガバナンスとの親和性に配慮し、アジャイルガバナンスの考え方をベースとしていることや、法的拘束力のないガイドラインとしていることが特徴である[17)]。アジャイルガバナンスは、政府、企業、個人・コミュニティーといったさまざまなステークホルダーが、自らの置かれた社会的状況を継続的に分析し、目指すゴールを設定した上で、それを実現するためのシステムや法規制、市場、インフラといったさまざまなガバナンスシステムをデザインし、その結果を対話に基づき継続的に評価し改善していくアプローチである。その運用から受ける評価を速やかに反映するだけでなく、より大きな外部状況変化に対する環境・リスク分析によるゴール自体の見直しも行う。

また、日本ディープラーニング協会に「AIガバナンスとその評価」研究会が発足し、そこでの検討に基づき、2021年7月に報告書「AIガバナンス・エコシステム―産業構造を考慮に入れたAIの信頼性確保に向けて―」が公開された。従来はAIガバナンスが1組織・1企業における内部ガバナンスの在り方という限定的な意味で用いられがちであったが、日本におけるAIサービスは、開発者、サービス提供者や運用者、利活用者などにわたるサプライチェーンが非常に長い構造を持つことから、組織を超えたガバナンスの仕組みを考えていくべきということが提言されている。

❷ –c. データ保護に関する法改正[18), 19)]

日本におけるデータ保護の法制度としては、まず2003年に公布、2005年に施行された、個人情報保護法を含む個人情報保護関連5法がある。その後、改正が議論され、改正個人情報保護法が2015年に公布され、2017年5月に施行された。この改正では、事業者間での轉転流通が認められる匿名加工情報の新設、要配慮個人情報の導入、高い独立性を持つ個人情報保護委員会の設立など、データを取得した個人の同意なしでデータを流通・利用活するための新しい枠組みが創設された。3年ごとの見直し規定も定められ、2020年に公布、2022年4月から施行でさらに改正が加えられた。この改正には、第三者への提供禁止請求・提供記録開示請求など本人の権利保護の強化、データの利活用の促進のため制約を緩和した仮名加工情報[10]の新設、その一方で法令違反に対する罰則を強化などが盛り込まれている。

10 仮名加工情報とは、他の情報と照合しない限り、特定の個人を識別できないように個人情報を加工したものである。加工によって一定の安全性を確保しつつ、匿名加工情報よりもデータの有用性を保ち、詳細な分析が可能になった。仮名加工情報を、他の情報と照合して、特定の個人を識別することは禁止される。

欧州（EU）では、1995年に制定されたデータ保護指令（Data Protection Directive: 95/46/EC）が存在したが、2012年からインターネット、デジタル化といった技術進化やグローバル環境変化を踏まえた全面的な見直しが進められた結果、パーソナルデータの取り扱い（Processing）と移転（Transfer）に関わる規則（Regulation）を定めた一般データ保護規則（General Data Protection Regulation：GDPR）が2016年4月に成立し、2018年5月から適用が開始された[11]。先のデータ保護指令は、各国に一定の法律の制定を義務付けているが、指令（Directive）が各国に直接適用されるわけではない。それに対してGDPRは、各国に直接適用されるため、運用面での位置付けが大きく異なる。規則の内容の面では、特にAIとの関係が深いものとして、プロファイリングに基づく自動意思決定に対する説明責任・透明性を要求していること（GDPR第22条）が挙げられる。その他にも「削除権」（「忘れられる権利」、同第17条）、「開示請求権」（同第15条）、「データポータビリティー権」（同第20条）、罰則の強化などにも特徴がある。

米国では、公的部門については1974年のプライバシー法が存在するものの、民間部門については包括法がなく、自主規制を基本としている。すなわち、企業が自ら公表しているプライバシーポリシーに違反した場合、公正取引委員会（Federal Trade Commission：FTC）がFTC法第5条「不公正または欺瞞的行為の禁止」に照らして取り締まる。規制対象を限定して個別領域ごとに個別法が制定されるセクトラル方式がとられている。ただし、Google、Meta、Amazon、AppleなどのBig Tech企業に膨大な利用者データが集まる状況に際して、2012年に「ネットワーク化された世界における消費者データプライバシー」という政策大綱が公表され、その中で、事業者によるインターネット上の追尾・追跡（トラッキング）を消費者が拒否（オプトアウト）できる「Do Not Track（DNT）」という概念を明確化して基本方針とした「消費者プライバシー権利章典」が提案された。また、「急変する時代の消費者プライバシー保護」レポートで、Privacy by Design、単純化した消費者の選択、透明性という3条件が枠組みとして勧告された。2015年には権利章典をもとにした「消費者プライバシー権利章典法案」が公開された。また、州法レベルでのさまざまなプライバシー保護法が制定されているが、特にカリフォルニア州は先進的な法律を制定してきた。例えば、2002年に同州が最初に制定したセキュリティー侵害通知法（California Security Breach Notification Act：情報漏洩が発生した場合に事後的な通知・報告を義務付け）は、その後、ほぼ全州が同様の法律を制定した。さらに、カリフォルニア州は2018年6月に新たに消費者プライバシー法（California Consumer Privacy Act：CCPA）を制定した（2020年1月施行）。CCPAはGDPRと同様にパーソナルデータの保護を強く打ち出しているが、GDPRと比べて、対象となる企業や個人の範囲は狭いものの、個人に付与される権利はより幅広い。

また、GDPRのようなデータ保護を強化する施策だけでなく、保護しつつもデータ利活用を進めやすくするような施策も考えられつつある。その一例として、2022年5月に公表された欧州ヘルスデータスペース規則案（European Health Data Space：EHDS）が挙げられる。現状は国ごとに取り扱いルールが異なっているヘルスデータについて、国内や国を超えて自分のヘルスデータの管理を可能にし、安全性を保ちつつ研究・イノベーション・公衆衛生・政策立案などへの活用を可能にすることが目指されている。これに対して、日本では、個人情報管理規定が政府・自治体・民間事業者ごとにバラバラで、活用が阻害されているという「2000個問題」や、医療分野の研究開発に資するための匿名加工医療情報に関する法律「次世代

11 GDPRでは、EU域内の個人に関するデータをEU域外へ移転することを原則として認めていないが、十分なデータ保護政策がとられている国であれば、十分性認定を受けることで、移転が認められる。日本は十分性認定を受けている。米国は認められていない。

医療基盤法」の見直し[12]など、課題が積み残されている。

❷ –d. AIに対応した知的財産戦略

国内では、内閣の知的財産戦略本部での議論の中で、AIに対応した知的財産戦略・法改正などが議論されている。2017年3月に関連する報告書[20]が公表された。

まず、機械学習の訓練（学習）データに関わる著作権法が改正され、機械学習のために、よりデータを利活用しやすくなった。もともと日本の著作権法は、コンピューターによる情報解析を目的とした複製などを許容する権利制限規定を有しており（旧47条の7）、営利目的であっても、第三者の著作物が含まれていても、一定限度で著作権者の許諾なく著作物を利用することが可能とされている。さらに、2019年1月1日施行の改正著作権法では、旧法では制限がかかっていたと解釈されるいくつかのケースに関して制限が緩和された（新30条の4第2号）。すなわち、旧法では訓練データを作成する主体と機械学習を実行する主体が同一であることを前提としていたのに対して、新法ではその前提が排除された。すなわち、作成した訓練データセットを他者に提供することも許容され、その利活用がいっそう促進される。

次に、訓練（学習）済みモデルの権利保護の課題について述べる。訓練済みモデルの再利用の仕方は、単純にモデルをそのままコピーして使うケース（複製）だけでなく、追加学習して使うケース（派生）、複数個の訓練済みモデルを組み合わせて使うケース（アンサンブル：Ensemble）、訓練済みモデルの振る舞い（どんな入力に対してどんな出力を出すか）の観測データを別のニューラルネットワークに学習させて新たなモデルを作るケース（蒸留：Distillation）が知られている[21]。このうちアンサンブルで使うモデルは複数個だが、一つ一つは複製モデルに相当するので、権利保護を考えるべきモデルの種類は複製モデル・派生モデル・蒸留モデルの3種類である。このうち、派生モデルと蒸留モデルは、もとの訓練済みモデルとの関係性の立証が難しいことが権利保護上の課題である。契約・特許権・著作権などでどこまで保護できるか、新しい権利による保護が必要か、営業秘密として不正競争防止法で保護できるケースはどのようなケースか、といった検討がなされている[20]。

また、AI生成物の著作権については、次のような解釈がされる[10]。人間がAI技術を道具として利用した創作物は、その人間に創作意図と創作的寄与があれば、その創作物は著作物であると認められる。一方、人間に創作的寄与が認められないケースは、AI創作物とされ、現行の著作権法では著作物と認められない。AIはパラメーターを少しずつ変えながら休むことなく膨大なバリエーションの生成物を出力することが可能なので、それらに著作権を与えたら、大きな弊害を生む。ただし、ここでいう創作的寄与というのがどの程度のものならば該当するのかは、今後の課題として残っている。加えて、AI創作物を人間による創作物だと偽られる懸念や、機械学習を用いた場合に生成物が訓練データと類似してしまう問題なども課題である。知的財産として新たな保護を与えるかは、そのメリットとデメリット、それが市場に与える影響などのバランスも考えておく必要がある[13]。2022年には、一見するとプロが描いたようなテイストの画像が簡単な説明文から生成できる画像生成AI（Text–to–Image）が、一般にも利用可能な形で提供され、大きな話題になった（「2.1.1 知覚・運動系のAI技術」参照）。この技術を用いた作品がアートコンテストで1位になったことなどもあり、アーティストやクリエーターからの反発も生じており、上述の課題が急速に顕在化している。

12 2022年12月27日に開催された次世代医療基盤法検討ワーキンググループ第7回（内閣府 健康・医療戦略推進事務局）では、医療情報の研究ニーズ、社会的便益の観点から、新たに「仮名加工医療情報」の作成・提供を可能とし、その際、個人情報の保護の観点から、仮名加工医療情報の提供は国が認定した利活用者に限定するという提案がなされている。

13 文献6）の11章「ロボット・AIと知的財産権」（福井健策）に詳しい。また、その中ではロボット・AIによるコンテンツ生成に伴って考えられるメリットとして、（1）大量化・低コスト化による知の豊富化、（2）テーラーメイドでの個別ニーズの汲み取り、（3）侵害発見・権利執行の容易化によるフリーライドの抑制、（4）新たな体験・発見・感動、その一方でリスク要因として、（1）価格破壊による創造サイクルの混乱、（2）知のセグメント化の進行・集合体験の欠落、（3）フリーライドの多発・プロセス複雑化による権利関係の混乱、（4）コピーの連鎖による知の縮小再生産、が挙げられている。

同様に文章生成AI（ChatGPTなど）を用いて、文学作品や論文・レポートを作ることも行われ始めている。ChatGPTは米国のMBA、法律、医療の試験に合格できるレベルにあるという報告もあり、論文や試験レポートへの使用を禁じる動きが既に出てきている（「2.1.2 言語・知識系のAI技術」参照）。

❸「社会におけるAI」のための技術開発

❶で導出した要件・目標の中には、その実現のために新たな技術開発が必要なものが含まれている。特に活発に取り組まれている技術課題として、機械学習における公平性や解釈性を確保するための技術開発や、プライバシー保護のための技術開発が挙げられる。公平性・解釈性に関する代表的な技術としては、公平性配慮データマイニング技術（Fairness-Aware Data Mining：FADM）や説明可能AI技術（Explainable AI：XAI）などの開発が進められている。FADMでは、グループ公平性・個人公平性などの公平性基準を定義し、それを用いて不公平さを検出する手法や、不公平を防止する手法が開発されており、XAIでは、深層学習のように精度が高いが解釈性が低いブラックボックス型モデルに近似的な説明を外付けする方式や、決定木や線形回帰のような解釈性は高くても精度に限界のあったホワイトボックス型モデルを場合分けなどによって精度を高める方式が開発されている（詳細は「2.1.4 AIソフトウェア工学」を参照）。プライバシー保護のための代表的な技術としては、解析対象データにおけるプライバシー保護のためのデータ匿名化技術、データベース問い合わせにおけるプライバシー保護のための差分プライバシー技術、計算過程におけるデータ内容の漏洩防止のための秘匿計算（秘密計算と呼ばれることもある）技術、プライバシーを保護しながらデータ分析を行うプライバシー保護データマイニング技術（Privacy-preserved Data Mining：PPDM）などがある（詳細は「2.4.3 データ・コンテンツのセキュリティー」を参照）。

❹ AIと社会との相互作用

AI技術は社会を変え、変わった社会がさらに発展あるいは安定化するために、また新たなAI技術を要求する、というような連鎖（AI技術→社会変化→AI技術→社会変化→…）が起こってくる。そのような社会変化やAI技術の発展から、人間の在り方や思考の仕方も影響を受ける。連鎖のスピードや方向に、必ずしも全ての人々が追従できるわけではない。連鎖がどのような方向へ進むかを迅速に予測・把握し、連鎖の進行をうまくコントロールするための対策を的確に講じていくことが望ましい。

社会への影響に関する話題の一例として、英国オックスフォード大学から2013年に発行された「雇用の未来（The Future of Employment）」と題する論文[22)]が挙げられる。「今後10～20年程度で、米国の総雇用者の約47%の仕事が自動化されるリスクが高い」という予想を示したことから、AI技術による職業や雇用機会の変化が盛んに論じられるようになった。さまざまなタスクで、特化型AIが人間を上回る精度・性能を示していることに加えて、大規模言語モデル・基盤モデルによって、人間によるものか判別困難な品質で多数のタスクに対応できるという汎用性の向上も示され（「2.1.2 言語・知識系のAI技術」参照）、AIの社会に与える影響が急激に拡大しつつある。このような状況から、「2.1.5 人・AI協働と意思決定支援」で述べるような、人・AI協働の在り方や、人の意思決定への影響を考えていくことや、「2.1.6 AI・データ駆動型問題解決」で述べるような、AI技術による社会・産業・科学の変革をより良い方向に進めていくことが重要になる。

また、人々がアクセスできる空間の広がりとして、メタバースや宇宙が注目されている。このような新たな活動空間において、自律性の高いAIや人間の能力を拡張するようなAIへのニーズは高く、AIの技術発展と社会の発展・拡大の相互作用を適切にコントロールしていくは、ますます重要な課題になっていく。

（4）注目動向

［新展開・技術トピックス］

❶ 欧州AI規制法案と米国AIリスク管理フレームワーク

［研究開発の動向］❷–aで述べたように、欧州と米国では、原則から実践へのトップダウン政策として、欧州AI規制法案（AI Act）と米国AIリスク管理フレームワーク（AI RMF）が発表された。ここでは、これら二つの内容を簡単に紹介する。

欧州AI規制法案では、AIをリスクの大きさによって、（a）容認できないリスク、（b）ハイリスク、（c）限定的なリスク、（d）最小限のリスク/リスクなし、という4段階に分類し、（a）に該当するAIは使用禁止とされ、（b）は事前に適合性評価、（c）は透明性の確保が必要とされる。違反すると、巨額の制裁金や欧州でのビジネスに制約がかかる可能性がある。

具体的にどのような応用例が該当するかというと、（a）には、潜在意識への操作、子供や精神障害者を相手とする搾取行為、社会的スコアの一般的な利用、公的空間での法執行目的の遠隔生体認証が挙げられている。（b）には、規制対象製品の安全要素（産業機械・医療機器など、法によって第三者認証の対象となるもの）や、特定分野のAIシステム（自然人の生体認証と分類、重要インフラの管理と運用、教育と職業訓練、雇用・労働者管理・自営業の機会、必須の民間サービスや公共サービス・利益へのアクセスや享受、法のエンフォースメント、移住・亡命および国境管理、司法運営と民主的プロセス）で人の安全や権利に影響を及ぼすリスクが高いものが対象となる。（c）には、自然人と相互作用するシステム（例えばチャットボット）、感情推定や生体情報に基づくカテゴリー形成を行うシステム、ディープフェイクなどが例として挙げられ、そのような仕掛けを使っていることを人に通知する透明性義務が課される。

2021年8月まで意見公募があり[14]、その後、修正案が検討されており、2023年中の発効が有力で、完全施行は最速で2024年後半と見られている。なお、この法案には域外適用条項があり、日本企業がEU域内で商品やサービスを提供する場合にも適用される。

一方、米国AI RMFは、2部構成になっており、第1部でAIに関わるリスクの考え方と信頼できるAIシステムの特徴を概説し、第2部でAIシステムのリスクに対処するための実務を説明している。この第2部では、具体的な対処の仕方を、マップ（リスクの特定）、測定（リスクの分析・評価など）、管理（リスクの優先順位付けやリスクへの対応）と、それらの統治（組織におけるリスク管理文化の醸成など）という四つの機能に分けて解説している。2022年3月に初期ドラフト、8月に第2ドラフトが公開され、2回の意見公募を経て、2023年1月に第1版（AI RMF 1.0）として発表された。

AI RMFは、欧州AI規制法案のような強い規制をかけるものではなく、AIシステムを設計・開発・導入・使用する者が自主的に参照できるものとされているが、米国NISTから発表され、業界への影響力が大きいため、今後、AI標準化の有力なベースとなっていくものと考えられている。

❷ パーソナルAIエージェント/サイバネティックアバターとELSI

各個人に関わるデータ（パーソナルデータ）は、サービス運営企業のところに、利用者データとして集められ管理される形態が大半であった。しかし、パーソナルデータの漏洩事故や利用者の意図せぬ利用などの懸念も生じ、パーソナルデータの管理を個人主導の形態へ移行させようという動きが進みつつある[18)]。自分のパーソナルデータがどこまでどの企業に開示されているのかを、自分自身で把握し、コントロールしたい（すべき）という考えである。そのために法律面では、欧州のGDPRのように、自己情報コントロール権（開示請求権、削除権・訂正権、データポータビリティー権などが含まれる）の確保が考えられている。

14 規制対象となるリスクの高いAIについての定義の明確化や、責任範囲の明確化を求める意見など、304件の意見書が寄せられたとのことである。日本からも経団連などから意見書が出された。

一方、技術・システム面では、自分の管理下にパーソナルデータを集約・管理するPDS（Personal Data Store）や、その管理を委託する情報銀行（情報信託銀行の略称）などの仕組みが考えられている。

しかし、このような個人主導のパーソナルデータ管理では、各個人の管理能力・情報リテラシーが低いとかえってリスクが高まる恐れがあることに加えて、管理すべき相手・情報量の増大や条件・関係の複雑化によって、人間には管理しきれないという状況も予想される。これに対して、前述のIEEE EAD1eでは、各個人とサービス運営企業（事業者）との間に入り、個人の代理として、事業者の提示するパーソナルデータの利用方法とサービスが各個人の決めた条件に合致するかどうかを判断し、事業者にパーソナルデータを渡してその事業者からサービスを受けるかどうかを決定する「パーソナルAIエージェント」（PAI Agent）の概念が導入された。

また、ムーンショット目標1において、身代わりとしてのロボットや3D映像などを示すアバターに加えて、人の身体的能力、認知能力および知覚能力を拡張するICT技術やロボット技術を含めた概念が「サイバネティックアバター」（Cybernetic Avatar：CA）と称されている。CA（具体的にはOriHimeのようなアバターロボット）を使うことで、身障者が遠隔で職業に従事するといった社会的な取り組みも行われている。 CAとそれを使う人間の関係は、1対1とは限らず、一人の人間が複数のCAを使うパターンもあれば、複数の人間で一つのCAを共同で操作するパターンもある。必ずしもCAの一挙一動を人間がコントロールするわけではなく、CAの振る舞いはある程度の自律化・自動化がなされたものになっていく。その意味で、PAI AgentとCAの役割はかなり近いものになる。

これらPAI AgentやCAと人間の関係は、これまでの社会にはなかった新しい様相をもたらし、ELSIの観点からさまざまな課題が生じるため、検討が進められている[11), 23), 24), 25), 26)]。例えば、PAI AgentやCAがそれを使う人間の意図通りに振る舞わないかもしれない。それで事故や問題が起きたときの責任の所在はどこにあるのか。CAを操る人間をじかに確認することができない状態で、CAに相対する人間はCAをどうすればトラストできるのか。人間、CA、PAI Agentの間でなりすましやのっとりが起きていないことはどうすれば確認できるのか。一人で同時に複数のCAを使ったり、一つのCAを複数人で共同操作したりするとき、人間の心的面にどのような影響が生じるのか。さまざまな面から分析や実験を進めていくことが望まれる。

また、個人のライフサイクルとPAI Agentが代理を果たすべき期間に関わる問題も考えておく必要がある。個人のパーソナルデータが発生し、それが存続する期間に対して、その個人が自分自身でそのデータを管理できる期間は限定される。胎児・幼児期にはパーソナルデータを自分で管理できないことはもちろん、身体的な死を迎える以前に認知症などによって自分では管理できなくなる可能性がある。さらに死後もパーソナルデータ管理（特に故人がSNS上に発信していた情報やデジタル的に管理していた情報、いわばデジタル遺産管理）の問題は残る。このような期間のパーソナルデータ管理について、PAI Agentに代理を委ねることが考えられるが、技術的な実現方法の問題だけでなく、法的な位置付け、プライバシーの扱い、代理の権限移譲の方法なども重要な課題であり、併せて検討してされている[23), 24), 26)]。

［注目すべき国内外のプロジェクト］

❶ GPAI（Global Partnership on AI）

前述の通り2019年5月に「人工知能に関するOECD原則」が出されたが、これを実践フェーズに移すために、2020年6月にGPAIが発足した。 GPAIは、人間中心の考え方に立ち、「人工知能に関するOECD原則」に基づき「責任あるAI」の開発・利用を、プロジェクトベースの取り組みで推進するために設立された、政府・国際機関・産業界・有識者などマルチステークホルダーによる国際連携イニシアチブである。2023年1月時点で、28カ国とEUが参加している。2022年11月に第3回年次総会が東京で開催され、同月より1年間、日本が議長国を務める。現在、「責任あるAI」「データガバナンス」「仕事の未来」「イノベーションと商業化」という四つのテーマについてワーキンググループが設置されており、専門家による

議論が行われている。日本からはこれら全ワーキンググループに専門家が参加している。

❷ 中国のAIによる社会監視システム

中国では、従来の「金盾」（Great Firewall）システムに加えて、「天網」（Sky Net）システムと「社会信用システム」の構築を進めており、政府による社会や国民の監視・管理が、AI技術を用いて強化されている。「金盾」はインターネット通信の検閲システムであり、ウェブ検索エンジンの検索語、電子メールやインスタントメッセンジャーの通信内容、ウェブサイトやSNS（Social Networking Service）のコンテンツなどに対して検閲・遮断が行われる。2003年頃から稼働し、その後も段階的に強化されている。「天網」は監視カメラネットワークで、2012年に北京市に本格的に導入され、2015年には中国内の都市エリアが100%カバーされた。2019年には監視カメラ27億台の規模になっている。顔認証技術が組み込まれており、人混みの中から指名手配犯を見つけ、逮捕できたという実績もあげている。深圳市では、交差点に設置された監視カメラから信号無視などの違反者を見つけ、警告する試みも実施された。

さらに中国政府は「社会信用システム」の構築計画（2014年～2020年）を発表した。所得・社会的ステータスなどの政府が保有するデータに加えて、インターネットや現実社会での行動履歴も含めて評価し、各国民の社会信用スコアを計算するという計画である。しかし、実際には、政府によるものではなく、民間の電子マネー運営企業・電子決済運営企業において、利用者のプロファイルや行動履歴から独自に信用スコアを算出し、そのスコアに応じて利用者に優遇や制限を与える（公共交通機関の割引・制限、病院診察やビザ取得手続きでの優遇など）ことが行われている。特にAlibaba（阿里巴巴）が展開する信用スコア「芝麻信用」は中国内で利用者が5億人を超えるという電子決済サービスAlipayと連動しており、大きな存在感を示している。

このようなさまざまな行動の監視や信用スコアに基づく賞罰によって、品行方正に振る舞う人々が増え、犯罪・違反の抑制や迅速な逮捕にもつながるという効果が得られているという。中国の「金盾」「天網」「社会信用システム」そのものは、表現の自由やプライバシーを重んじる欧米・日本には適合しないシステムであるが、AIが組み込まれた社会の一形態として非常に興味深い。なお、中国も参加しているユネスコの第41回総会で2021年11月に満場一致で採択された「AI倫理勧告」では、人権を守り、社会監視や社会的格付けのためにAIを使用すべきでないとしており、中国はここに挙げた社会監視システムを今後どうしていくのか注目される。

（5）科学技術的課題

❶「社会におけるAI」の課題抽出・目標設定に関わる研究開発課題

AI社会原則・AI倫理指針がさまざまな国・機関から出され、国際的な議論もなされたことから、高い抽象度で記述される原則のレベルにおいては、世界共通の意識が持たれつつある。しかし、より具体化された場面、細則においては、国・地域固有の文化や社会の価値観には違いが表れる。その一例を示したのが、米国マサチューセッツ工科大学（MIT）メディアラボのモラルマシン実験である。これは自動運転車版のトロッコ問題（倫理的ジレンマ）に関する思考実験で、自動運転車にブレーキ故障などが生じ、事故で犠牲者が出ることが避けられない状況を示し、その中で一部の人だけ免れるとしたとき誰が優先されるべきかを、さまざまな人々に問い、233の国・地域、230万人からのべ4000万件の回答を得た。2018年10月に発表された分析結果[27)]によると、例えば、個人主義的な文化を持つ地域では、より多くの人数を救うことが優先される傾向や、一人当たりのGDPが低く、法の規律も低い地域では、法順守違反に寛容だという傾向や、経済格差の大きい地域では、社会的な地位の高さが優先される傾向などが見られ、地域の文化的背景との相関、各地域の特徴、地域間の類似性などが示された。

ここで示されたような倫理観の多様性に対して、倫理ルール作りをどのように進めていくべきか、あるいは、ある倫理観にしたがって作られた製品・サービスの地域ごとの受容性をどう考えていくかなど、より検

討を深めていくことが望まれる。そのためには、技術者・利用者・政策関係者・企業経営者など、さまざまな立場の視点を盛り込み、具体的な問いを立てて論じていくことが有効と思われる[1)]。デザイン、アートの分野で注目されているスペキュラティヴ・デザイン（問いを立てるデザイン）[28), 29)]の取り組み・考え方も参考になる。また、AIの軍事利用問題もAI倫理に関わる重要課題であるが[4)]、前述のAI社会原則・AI倫理指針において、EADv2以外では踏み込んだ記載は見られない。自国第一主義が台頭してきている世界情勢の中で、重要だが取り組み方の難しい問題である。軍事利用の観点で一番の懸念と思われるのは、人間の関与なしに自律的に攻撃目標を設定することができ、致死性を有する自律型致死兵器システム（Lethal Autonomous Weapons Systems：LAWS）である。これについては、特定通常兵器使用禁止制限条約（CCW）の枠組みに基づき、CCW締結国の中で2014年から会合が持たれ、2019年11月には11項目から成る「LAWSに関する指針」が示された[15]。引き続き、LAWSの定義（特徴）、人間の関与の在り方、国際人道法との関係、既存の兵器との関係など、規制の在り方が主要論点とされており、規制に対する推進派・穏健派・反対派に各国の立場が分かれている。ロシアによるウクライナ侵攻をはじめ国際的対立が先鋭化する状況において、ますます重要でありながら、合意の難しい問題になりつつある。

❷「社会におけるAI」のための制度設計に関わる研究開発課題

研究開発の動向や注目動向のパートで取り上げたように、一部は日本としての戦略・制度改革方針に沿って手が打たれつつあるが、制度設計が追い付いていない課題も多く残され、新たな課題も生まれており、引き続き検討・施策推進が求められる。特にAIにはブラックボックス性があるため、予見可能性に基づく過失責任主義（ハザードベース規制）に馴染みにくい。そこで、厳格責任を含むリスクベース規制の考え方を取り入れることや、大きな罰則・制裁を加えるより操作や原因究明への協力を促進する訴追延期合意制度（Deferred Prosecution Agreements：DPA）の適用などが考えられている。

また、AIエージェントとサイバネティックアバターとロボットの間の境界は薄れつつあり、制度設計ではそれらを合わせて検討していく必要があろう[10), 30)]。

国の制度設計においては、他国の動きもウォッチし、国際的な方向性との整合性・連動性も考慮していくことも必要である。欧州は各個人の権利を重視し、法制度でAIをコントロールしようとする傾向が見られる。米国はAI技術がもたらすベネフィットとリスクのバランスを法制度で調整しようとしている[16]。中国は［注目すべき国内外のプロジェクト］❷に示したように、AI技術と法制度を用いて、国による監視・管理を強めている。

このような中、欧州がGDPRやAI規制法案などハードローに踏み込んで、国際的ルールメイキングを先導している。日本は国際的ルールメイキングに関わる人材が限定的であることが大きな課題であり、また、人権や正義に根差した理念重視の議論を行う欧州に対して、日本は社会的価値観の曖昧さや議論を避けがちな国民性から受け身の対応になりがちである。そのような状況の中でもGPAIやISO/IEC JTC 1/SC 42国際標準化活動などでは健闘していると言える。そこでも見られるが、理念から論ずるよりも、むしろ、具体的なケースから実践的なルール作りやトラスト形成を積み上げるというのが日本らしいアプローチかもしれない。

15 外務省「自律型致死兵器システム（LAWS）について」（2020年11月4日）
https://www.mofa.go.jp/mofaj/dns/ca/page24_001191.html（accessed 2023-02-01）

16 米国ホワイトハウスによるAIガイドライン「Guidance for Regulation of Artificial Intelligence Applications」（Memorandum for the Heads of Executive Departments and Agencies）では、重要な価値・権利を保護することが大切だとしつつも、過度の規制によってイノベーションの促進が阻害されることは避けるという考えが示されている。

❸「社会におけるAI」のための技術開発に関わる研究開発課題

既に技術開発が推進されているプライバシー保護技術や機械学習の公平性・解釈性を確保する技術に関しては、「2.4.3 データ・コンテンツのセキュリティー」「2.1.4 AIソフトウェア工学」の節で今後の研究開発の方向性や課題を述べている。システム開発の観点では、従来のITリスクだけでなくAI ELSI面も含んだ、より複雑なAIリスクを考えていくべきということが指摘されている[31]。

また、❷と❸の両方に関わる課題、すなわち、制度設計と技術開発の両面から取り組む必要がある課題として、社会的なトラスト形成[32), 33)]が挙げられる。AIのブラックボックス問題に対して、説明可能AI（XAI）の技術開発は重要であるが、説明は近似なので、そこから外れる現象はどうしても残る（公平性を偽装するFairwashing[34)]も可能だと指摘されている）。高度化するフェイク生成を見破るために技術的なアプローチは不可欠であるが、全てを見破ることができるわけではない。技術開発だけでは限界があり、技術開発による対策と制度設計の適切な組み合わせによるトラスト形成が重要になる。断片的な情報だけから判断したり、ある一面だけを見て信じ込んだりすることはとても危うく、多面的・複合的な検証・保証で支えていくことが必要になる。また、ある程度の時間をかけて事例・実績が蓄積されたり、万が一のケースが保険などで補償されたりといったことを通してトラストが形成され、社会に受容されていくという側面もある。その具体的な取り組みや動向については「2.4.7 社会におけるトラスト」に記載しているので参照いただきたい。

❹ AIと社会との相互作用に関わる研究開発課題

AI技術発展と社会変化のスパイラルを見据え、そのようなスパイラルを組み入れた社会システムの発展プロセスのモデル化や社会システム設計の方法論の研究開発にも取り組んでいく必要がある。その中で、上記の❶「社会におけるAI」の課題抽出・目標設定、❷「社会におけるAI」のための制度設計、❸「社会におけるAI」のための技術開発を有機的に連携させていくことが重要である。また、職業の変化や人材育成・教育への取り組みも、その中で描いていくことが必要である。

（6）その他の課題

❶ RRIの推進と支援体制

「社会におけるAI」への取り組みではRRI（Responsible Research and Innovation：責任ある研究・イノベーション）の考え方が不可欠である。RRIとは、新しい技術の創出・展開を進めるにあたって、生み出される成果が倫理的・法的・社会的に受容可能で、社会的価値・持続可能性などの面でも好ましいものであることを担保しながら取り組むことである。 RRIは2000年代前半から欧米で議論されており、欧州のHorizon 2020においても政策的課題として重視されている。 RRIを成り立たせる要件として、予見的であること（Anticipatory）、応答的であること（Responsive）、熟議的であること（Deliberative）、自己反省的であること（Reflective）が挙げられている[14), 35)]。

ELSIやRRIに関する研究者の取り組みを強化するためには、個々の研究者に自覚を持たせるための継続的な教育・啓発とともに、事前検証と事後対応の両面で相談機会を促進し、ELSI問題を適切に解決するための仕組み作りも重要である。つまり、ELSIガイドラインなどを設定して研究者本人に任せるというだけでなく、知的財産センターが研究者の知的財産の創造・保護・活用を多面的に支援・促進するのと同じような、組織的な支援体制作りが求められる。

❷ 多様な視点・考え方の取り込みと具体化に基づく議論

AI・ロボティクスに関わる情報科学・工学分野の研究者・技術者だけでなく、倫理学者、哲学者、法学者、憲法学者、社会学者、政治学者、経済学者などの人文社会科学の研究者も検討に参画し、また、研究者・技術者だけでなく、利用者、政策関係者、企業経営者など、さまざまなステークホルダーの視点も

取り込んで議論・検討を進めていくことが望ましい。その際、さまざまな人々の間で、高い抽象度の総論で意見の一致を見るというレベルにとどまらず、具体化された問題・シーンに踏み込んで議論を深めることが、実施における具体的な問題が何か、施策として何が足りないかなどを見極めるために効果的である。深い議論につながるような問いを立てることや特区制度の活用なども今後重要になってくると思われる。

（7）国際比較

❶「社会におけるAI」の課題抽出・目標設定、❷「社会におけるAI」のための制度設計、❸「社会におけるAI」のための技術開発、❹AIと社会との相互作用 という四つの活動のうち、❶❷を基礎研究、❸❹を応用研究・開発として扱い、下表にまとめる。

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	◎	→	人工知能学会、内閣府、総務省、経済産業省などによる学会・政府主導のガイドライン策定が推進されてきた。内閣府「人間中心のAI社会原則」を発信し、G20などでの国際的議論に反映させた。データ保護では、事業者間での轃転流通が認められる匿名加工情報の新設などを盛り込んだ改正個人情報保護法を施行（その後、仮名加工情報も導入）。
	応用研究・開発	○	↗	プライバシー保護技術やPDS・情報銀行への取り組みに加えて、パーソナルAIエージェントの検討なども進められている。
米国	基礎研究	◎	→	産業界主導で取り組み（FLI、Open AI、Partnership on AIなど）が始まり、それを追うように学術界での取り組み（AI 100、IEEE EADなど）が立ち上がった。 EADはIEEE標準化を並行して進めており、NIST AI RMFを含め、国際的な影響力が大きい。
	応用研究・開発	◎	→	プライバシー保護技術に関わる理論的アイデア、機械学習の解釈性に関するXAI研究など、米国の大学・企業の研究者から提案され、実装への取り組みも活発で、研究者層も厚い。
欧州	基礎研究	◎	→	オックスフォード大学のFHI、ケンブリッジ大学のCSERなど、英国の大学・研究機関を中心に、比較的早い時期から取り組まれており、Ethics Guidelines for Trustworthyからさらに AI規制法案などのハードロー化を推進。データ保護では、AI処理の説明責任・透明性の要求や忘れられる権利などを盛り込んだ一般データ保護規則GDPRを施行。
	応用研究・開発	◎	→	プライバシー保護の基礎研究や自動運転の制度設計への取り組みで実績がある。
中国	基礎研究	△	→	個人情報保護も含む中国インターネット安全法（中華人民共和国網絡安全法）の制定、次世代AIガバナンス原則などの公表が行われたものの、その実践面においては顕著な進展は見られない。
	応用研究・開発	△	↗	倫理・プライバシー保護面の取り組みは弱いが、AI技術開発とその社会実装への取り組みは急成長している。中国独自のAI監視・管理社会のための技術開発・システム化も注目される。
韓国	基礎研究	△	→	ロボットと人間との関係について定めたロボット倫理憲章の草案が2007年に産業資源部によって発表された。
	応用研究・開発	△	→	特に目立った活動は見られない。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDSの調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

参考文献

1）江間有沙,『AI社会の歩き方：人工知能とどう付き合うか』(化学同人, 2019年）.
2）Mark Coeckelbergh, *AI Ethics*（The MIT Press, 2020）.（邦訳：直江清隆訳,『AIの倫理学』, 丸善出版, 2020年）
3）保科学世・鈴木博和,『責任あるAI：「AI倫理」戦略ハンドブック』(東洋経済新報社, 2021年）.
4）中川裕志,『裏側から視るAI：脅威・歴史・倫理』(近代科学社, 2019年）.
5）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」, CRDS-FY2018-SP-03（2018年12月）.
6）Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*（PublicAffairs, 2019）.（邦訳：野中香方子訳,『監視資本主義: 人類の未来を賭けた闘い』, 東洋経済新報社, 2021年）
7）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：複雑社会における意思決定・合意形成を支える情報科学技術」, CRDS-FY2017-SP-03（2018年12月）.
8）松尾豊・他,「人工知能と倫理」,『人工知能』(人工知能学会誌）31巻5号（2016年9月）, pp. 635-641.
9）江間有沙,「「人工知能と未来」プロジェクトから見る現在の課題」,『人工知能学会第29回全国大会』2I5-OS-17b-1（2015年）. DOI: 10.11517/pjsai.JSAI2015.0_2I5OS17b1
10）弥永真生・宍戸常寿（編）,『ロボット・AIと法』(有斐閣, 2018年）.
11）中川裕志,「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』(総務省学術雑誌）3巻2号（2020年）, pp. I-1-23. DOI: 10.24798/jicp.3.2_1
12）福島俊一,「AI品質保証にかかわる国内外の取り組み動向」,『情報処理』(情報処理学会誌）63巻11号（2022年11月）, pp. e1-e6.
13）中川裕志,「デジタル社会におけるAIガバナンス―倫理と法制度―」,『情報処理』(情報処理学会誌）62巻6号（2021年6月）, pp. e34-e39.
14）江間有沙,「倫理的に調和した場の設計：責任ある研究・イノベーション実践例として」,『人工知能』(人工知能学会誌）32巻5号（2017年9月）, pp. 694-700.
15）デロイトトーマツコンサルティング合同会社,「AIのガバナンスに関する動向調査 最終報告書（公開版）」, 令和元年度内外一体の経済成長戦略構築にかかる国際経済調査事業（2020年）.
16）情報処理推進機構AI白書編集委員会（編）,「制度改革（国内）」,『AI白書2022』(KADOKAWA,2022年）, pp. 368-396（4.3節）.
17）橘均憲,「「人間のためのAI（human-centric AI）」を実現する社会実装の道筋 ～AI 社会原則とAIガバナンス・ガイドライン～」,『情報処理』(情報処理学会誌）63巻9号（2022年9月）, pp. e1-e7.
18）中川裕志・他,「特集：パーソナルデータの利活用における技術および各国法制度の動向」,『情報処理』(情報処理学会誌）55巻12号（2014年11月）, pp. 1333-1380.
19）総務省,『情報通信白書（令和4年版）』(2022年）.
20）知的財産戦略本部 検証・評価・企画委員会 新たな情報財検討委員会,「新たな情報財検討委員会報告書 －データ・人工知能（AI）の利活用促進による産業競争力強化の基盤となる知財システムの構築に向けて－」(2017年3月）.
21）丸山宏・城戸隆,「機械学習工学へのいざない」,『人工知能』(人工知能学会誌）33巻2号（2018年3月）, pp. 124-131.
22）Carl Benedikt Frey and Michael A. Osborne, “The Future of Employment: How Susceptible are Jobs to Computerisation?”,（September 17, 2013）. https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

（accessed 2023-02-01）
23）加藤綾子・中川裕志,「パーソナルAIエージェントの社会制度的位置づけ」,『電子化知的財産・社会基盤（EIP）研究報告』2020-EIP-90（25）（2020年11月18日）.
24）中川裕志,「ディジタル遺産のパーソナルAIエージェントへの委任」,『電子化知的財産・社会基盤（EIP）研究報告』2020-EIP-90（26）（2020年11月18日）.
25）中川裕志,「AIエージェント、サイバネティック・アバター、自然人の間のトラスト」,『情報通信政策研究』（総務省学術雑誌）6巻1号（2020年）, pp. IA-45-60. DOI: 10.24798/jicp.6.1_45
26）Hiroshi Nakagawa and Akiko Orita, “Using deceased people’s personal data”, *AI & Society*（2022）. DOI: 10.1007/s00146-022-01549-1
27）Edmond Awad, et al., “The Moral Machine experiment”, *Nature* Vol. 563（2018）, pp. 59-64. DOI: 10.1038/s41586-018-0637-6
28）Anthony Dunne and Fiona Raby, *Speculative Everything: Design, Fiction, and Social Dreaming*（The MIT Press, 2013）.（邦訳：久保田晃弘・千葉敏生訳,『スペキュラティヴ・デザイン：問題解決から、問題提起へ。―未来を思索するためにデザインができること』, ビー・エヌ・エヌ新社, 2015年）
29）長谷川愛,『20XX年の革命家になるには―スペキュラティヴ・デザインの授業』（ビー・エヌ・エヌ新社, 2020年）.
30）Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts*（Springer, 2013）.（邦訳：新保史生・松尾剛行・工藤郁子・赤坂亮太訳,『ロボット法』, 勁草書房, 2018年）
31）中島震,『AIリスク・マネジメント：信頼できる機械学習ソフトウェアへの工学的方法論』（丸善出版, 2022年）.
32）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：デジタル社会における新たなトラスト形成」, CRDS-FY2022-SP-03（2022年9月）.
33）科学技術振興機構 研究開発戦略センター,「俯瞰セミナー&ワークショップ報告書：トラスト研究の潮流～人文・社会科学から人工知能、医療まで～」, CRDS-FY2021-WR-05（2022年2月）.
34）Ulrich Aivodji, et al., “Fairwashing: the risk of rationalization”, *Proceedings of the 36th International Conference on Machine Learning*（ICML 2019; June 9-15, 2019）, PMLR 97: pp. 161-170.
35）平川秀幸,「責任ある研究・イノベーションの考え方と国内外の動向」, 文部科学省 安全・安心科学技術及び社会連携委員会（第7回）資料4-3（2015年4月14日）.