Hitotsubashi University
Institute of Innovation Research

AI原則の体系化と今後のガバナンスの方向
～デジタル・AIにおけるイノベーションと社会制度の共進化～

市川類

IIR Working paper WP#20-15

2020年10月

Structuring AI Principles for the Future of AI Governance – Co-evolution of Innovation and Institutions in Digital and AI Technologies.

Tagui Ichikawa

一橋大学イノベーション研究センター

東京都国立市中2-1
http://www.iir.hit-u.ac.jp

AI原則の体系化と今後のガバナンスの方向
～デジタル・AIにおけるイノベーションと社会制度の共進化～

一橋大学イノベーション研究センター
市川類
2020年10月2日

１．　問題意識

イノベーションは、社会に対して新たな価値を創造するだけでなく、これまでの価値に対し破壊をもたらす。近年開発・普及が進みつつある人工知能（AI）技術は、破壊的イノベーションの一つであり、その破壊性がゆえに、既存の産業・雇用構造（格差を含む）だけでなく、社会の有する価値観などにも大きな影響を与えるとされる。

このような中、近年、世界の各国政府・機関において、多くの「AI原則」が策定されてきている。これらは、「AI倫理」とも言われ、このような社会の有する価値観に影響を与える可能性のあるAIに対して、ある種のガバナンスをかけようとする動きであると解釈される。また、これらのAI原則をベースに、AIに対する具体的な規制・ガバナンス構造を構築、制度化していこうとする動きも一部に見られる。

しかしながら、これらの「AI原則」「AI倫理」は、基本的人権の尊重、公平性、透明性、セキュリティ、説明責任といった、ある意味で当たり前の各項目の羅列にしか過ぎない。このため、AIの開発・利用者におけるチェックリストとしては有効かもしれないが、政府において今後のガバナンス体制の在り方を検討する立場からみると、それぞれの項目がAI技術のどのような特徴に基づき何故原則の一つとして提示されているのか、また、結局どのような規制・ガバナンス構造を構築することが求められているのかについて、整理されておらず、その結果、今後のAI技術・イノベーションの進化に伴い、AIに対して、何故、どのようなガバナンスが必要とされるのかについて、共通の認識が醸成されていない状況にある。

このような問題意識のもと、本論文では、世界各国で策定されているAI原則のレビューを行った上で、イノベーションと社会制度（ガバナンス）の共進化という枠組みの下で、これらのAI原則の分析・体系化を行い、それを基に、今後のAIに係るガバナンスの方向及びその限界について考察を行う。

２．AI原則を巡る世界の動き

（１） 世界におけるAI原則を巡る動き

＜世界のAI原則＞

ハードウェア・ネットワークなどの先端情報技術の継続的なイノベーションの進展に伴い、近年、デジタル・AI技術は大きく進展しつつある。特に最近の深層学習技術のブレークスルーにより、デジタル技術の機能・能力は急速に拡充し、今後の世界の経済成長のエンジンとしての期待が高まる一方で、他の技術にもまして、社会の有する価値観に対し大きな影響を及ぼす可能性のある技術として認識されつつある。

このような認識の下、世界の各国政府等において、AI戦略が打ち出されてきているが、その一環として、いわゆるAIのELSI（倫理的法的社会的課題）的側面への対応が、その重要な項目の一つとして打ち出されてきている。

その対応の一環として、世界の各国政府・国際機関において、多くの「AI原則」が策定されつつある。もともと、国・国際機関としては、日本（総務省）の動きが世界的に早く、その後、欧州、カナダ（大学）、中国などがAI原則を策定し、それらの各国の動きを踏まえた形で、2019年5月にOECDのAI原則（「信頼できるAIのための責任あるスチュワードシップに係る原則」）が、また、翌月ほぼそれと同内容のG20のAI原則が、米国を含む国際的な主要国が合意した形で策定されている。（各AI原則を巡る詳細な経緯は、参考１を参照。）

図表１：世界各国・機関のＡＩ原則を巡る経緯[1]

国等	機関	概要
日本	総務省IICP	• 2015年1月研究会開催発表、同6月報告（原則の検討を提言） • 2016年1月検討会議開催発表、同4月中間報告（G7に8原則案を提示）、同6月報告（同8原則を記載） • 2017年7月「国際的な議論のためのAI開発ガイドライン案」（9原則）発表 • 2018年7月「AI利活用原則案」発表、2019年8月「AI利活用ガイドライン」（10原則）発表
	内閣府	• 2016年5月「AIと人間社会懇談会」開始、2017年1月報告書発表 • 2018年5月「人間中心AI社会原則検討会議」開始、同12月「案」発表、2019年3月「同原則」発表
	（人工知能学会）	• 2014年12月倫理委員会設立、2015年6月公開討論会、2016年6月「案」公開、2017年2月「倫理指針」発表。
欧州	EC	• 2018年3月、EGE「人工知能、ロボット及び自立システムに係る声明」発表。原則策定提案（9原則例示） • 2018年5月AI-HLEG設置、同年12月「ドラフト」発表、2019年4月「信頼できるAIに向けた倫理ガイドライン」発表。同時に「人間中心AIにおける構築」発表、評価リストの試験・フィードバックを開始 • 2020年2年「AI白書」発表。新法による法規制を志向
米国（民間含む）	ホワイトハウス	• オバマ政権時：2016年5月検討開始、同年10月「AIに未来に向けた準備」発表（公正、安全、ガバナンス） • トランプ政権時：2019年2月米国AIイニシアティブ署名、2020年1月「AI応用の規制ガイダンス」発表。非規制的アプローチを志向
	（FLI）	• 2014年3月創設。2015年1月、公開質問状公開、2017年1月「アシロマ原則」発表（23の原則）
	（PAI）	• 2016年9月創設。Tenetを発表
	（IEEE）	• 2016年12月「倫理的に揃えられたデザイン」ver1発表(4原則)、2017年12月同ver2(5原則)、2019年6月同First Edition（8原則）を承認・発表
カナダ	（モントリオール大学）	• 2017年11月「AIの責任ある開発に関する宣言（第一版）」発表、2018年12月「AIの責任ある開発に関するモントリオール宣言」発表（10原則）
中国		• 2019年3月、国会次世代AIガバナンス専門委員会設置、同年6月「次世代AIガバナンス原則－責任を有するAIの発展」を発表（8原則）
シンガポール		• 2018年6月「AIガバナンスと倫理イニシアティブ」立ち上げ（2原則） • 2019年1月モデルAIガバナンス枠組第一版発表 • 2020年1月同第二版及び組織実行・自己評価ガイド（ISAGO）を発表

[1] 出典：「参考１」より筆者作成。
機関の項において、括弧で閉じているものは、政府ではなく、非営利機関等によるもの。

OECD/ G20		• 2018年5月、CDEPはAIGO創設に合意 • 2019年5月、閣僚会議にて「AIに係る理事会勧告（信頼できるAIのための責任あるスチュワードシップに係る原則）」を採択 • 2016年6月、G20貿易・デジタル経済大臣会合、首脳会合にて、「G20AI原則」を発表。
UNESCO		• 2020年5月、AI倫理に関する勧告一次案発表

なお、本ワーキングペーパーで議論の対象とするAI原則とは、具体的には、図表２の通り。

図表２：本ワーキングペーパーで対象とするAI原則

国・機関	AI原則
日本（総務省）	「**国際的な議論のためのAI開発ガイドライン案**」（2017年6月）の開発原則（9原則）、「**AI利活用ガイドライン**」（2019年8月）の利活用原則（10原則）
日本（内閣府）	「**人間中心AI社会原則**」（2019年3月）のAI社会原則（7原則）
欧州（EC AI-HLEG）	「**信頼できるAIに向けた倫理ガイドライン**」（2019年4月）の「信頼できるAIの要件（7要件）」
FLI（非政府）	「**アシロマ原則**」（2017年1月:23原則）特に「倫理・価値」に係る13原則
IEEE（非政府）	「**倫理的に整列されたデザイン**－自律的・知的システムでの人類の幸福への重点化に向けたビジョン」（2019年6月）の「一般原則（8原則）」
モントリオール大学（非政府）	「**AIの責任ある開発に向けたモントリオール宣言**」（2018年12月:10原則）
中国	「**次世代AIガバナンス原則**」（2019年6月）（8原則）
OECD	「AIに関する理事会勧告」（2019年5月）のうち、「**信頼できるAIのための責任あるスチュワードシップに係る原則**」（5原則）
G20	「G20 AI原則」（2019年6月）のうち、同上部分（5原則）
UNESCO	「**AIの倫理に関する勧告**（ドラフト第一版）」（2020年5月）の「原則」部分（グループ１、グループ２）（13原則）

これらの世界各国の「AI原則」は、一般的には、AIを開発・利用する企業等[2]に対して、その開発・利用にあたって留意すべき事項を5～10項目強程度に類型化して、記載したものであり、非拘束の制度である。概ねの内容は、基本的人権の尊重、公平性の確保、説明責任、透明性・説明可能性の確保、プライバシーの確保、セキュリティの確保などの項目とそれぞれに係る留意事項の説明になっている。

ただし、その項目の分類・整理方法は、それぞれ全く異なっており、おそらくAI原則の数だけその整理・分類方法があるという感じになっている。なお、内容や視点についても、それぞれのAI原則によって若干異なり、例えば、OECDのAI原則は、比較的高いレベルの規範を記載しているのに対し、ECのAI原則は、厳密には「要件」であるため、具体的に取り組むべき課題について、ある程度踏み込んでいるなどの特徴がある。

[2] OECDのAI原則では、「AI Actorは、～すべきである」という記載が中心。AI actorsとは、AIのライフサイクル（設計～検証～導入～運用・監視）において積極的役割を果たす個人・組織と定義。
ECのAI原則では、「AI Systemは、～されるべきである」という記載が中心であるが、一部、「AI Practitionerは、～すべきである」という記載を含む。ここで、AI practitionersとは、AIシステムを開発、導入、利用する個人又は組織（最終利用者、消費者を除く）。

これらのAI原則については、これまでに、上述のような世界各国のAI原則を比較し、それぞれの項目の内容を紹介するような論文[3]や、これらのAI原則を踏まえて今後法制化の検討をすべきとする論文[4]などは存在する。

しかしながら、これらの「AI原則」については、現在進行中の取組であり、AI原則に含まれるような各項目が、AI技術のどの特徴に基づき、それぞれどのような社会課題に対応するために必要なのかなど、今後のAIガバナンスの体系の検討に必要な十分な分析や体系化立てた説明がなされていないのが現状である。

＜本ワーキングペーパーでのAIの定義＞

本ワーキングペーパーにおいて、「AI」技術とは、機械学習技術を含め、高度で多量のデータ処理を行うことにより、一定の自律的判断を行う「知的な」ソフトウェア技術の一種として概ね捉えるものとする。なお、「AI」の定義については、ISO/IEC JTC1 SC42などで議論は行われているが、現時点で必ずしも明確な定義は決定されていない。

近年、デジタル技術（ソフトウェア技術・ハードウェア技術）が高度化する中で、機械学習技術をはじめとするこのような「AI」技術は、当該デジタル技術から構成される各種の情報システムの中に主要かつ不可欠な一部として組み込まれつつあり、IoT、ロボットの活用も含め、一定の自律的判断を行う「知的な」情報システムとして機能することになる。

このため、本ワーキングペーパーにおいては、「IoT、ロボットの活用も含め一連のデジタル・AI技術から構成され、認識から判断までに係る一定の自律的判断を行う高度な情報システム」として「AIシステム」を定義し、「AIガバナンス」の議論の対象として、使用することとする。したがって、便宜上「AI」という用語は利用しているものの、むしろ「デジタル・AI技術を利用した高度な情報システム」として捉えるべきものであることに留意することが必要である。

なお、世界各国のAI原則におけるAIの定義を、参考２に示す。これらの定義も各種様々であるが、いずれの場合も、狭義の「AI」技術を含む、情報システム全体を議論の対象としている。

（２）AI原則を巡る問題意識の変化と今後の方向

＜何故AI原則なのか？：AI原則の策定に係る技術・社会認識の変化＞

それでは、そもそも何故、世界各国において、このようなAI原則に取り組もうという動きが出てきたのであろうか。

これらのAI原則は、一般的に、AIに対してガバナンスをかけようとする制度的な取組の一種ではあるが、いわゆる非拘束の制度であり、通常の規制のように特定の義務を課すもので

---

[3] 中川 裕志（理化学研究所）「AI 倫理指針の動向とパーソナル AI エージェント」総務省 学術雑誌『情報通信政策研究』 第３巻第２号（2020年3月30日）
https://www.soumu.go.jp/main_content/000679318.pdf
https://www.soumu.go.jp/iicp/journal/journal_03-02.html

[4] 新保 史生（慶應義塾大学）「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」総務省 学術雑誌『情報通信政策研究』 第３巻第２号（2020年3月30日）
https://www.soumu.go.jp/main_content/000679321.pdf
https://www.soumu.go.jp/iicp/journal/journal_03-02.html

はない。このような制度としての「AI原則」を策定しようとする動きは、AI技術の以下の二つの特徴によるものと整理できる。

- AI技術は、汎用的技術であり、かつ、破壊的イノベーションを引き起こす技術であること。そのため、負の側面も含めて、社会に幅広くかつ深い大きな影響を与えるものであること。したがって、何らかのガバナンスが求められていること。
- 一方、現時点でAI技術・イノベーションの将来は不透明であること。AI技術は、まだ普及の初期過程にあり、社会に対する今後の具体的な影響が現時点で特定困難であり、したがって、現時点で具体的な規制を設けることが困難であること。

特に後者（将来に対する不透明性）の観点に関しては、実際に、世界各国において、AI原則を策定した背景・経緯を見ると、この数年においても、社会による技術に対する認識によって変化が生じてきていることが読み取れる。すなわち、当初は、特に、汎用人工知能（AGい、強いAI）やシンギュラリティ（技術的特異点）に係る議論など、技術の将来に係る不透明性や国民市民の不安・恐怖感を含めた社会認識によって突き動かされていたのに対し、その後は、むしろ現実的にAIに関して問題提起されているような各種課題に対して、包括的な対応が必要と認識されるようになってきている[5]。

以下においては、その世界におけるAI原則の策定を巡る経緯について簡単に整理をする。

＜AGIに対する懸念から現実の懸念への対応へ＞

AI原則の必要性の議論が開始されたのは、2015年頃である。この時期は、ディープラーニングの普及が始まり、人工知能に対して急激に関心が高まる[6]一方、シンギュラリティ（技術的特異点）などの議論など汎用人工知能（AGI、強いAI）に対する国民の不安（いわゆるAI脅威論）があった時期となる。

そのような中、世界各国の政府機関の中で、AI原則論の必要性を初めて唱えたのは、（筆者の知る限り）日本・総務省である。総務省の情報通信政策研究所（IICP）は、2015年1月に、研究会の開催を発表しているが、その際、問題意識としては、シンギュラリティを含め、長期的な（SFのような）世界を想定した記述となっている。具体的には、

> *「2045年にはコンピュータの能力が人間を超え、技術開発と進化の主役が人間からコンピュータに移る特異点(シンギュラリティ)に達するとも議論されるなど、その処理能力は加速度的に高まっています。」*
>
> *「ビッグデータ、人工知能、ロボット等を通じて、既に私たちはこれら技術の恩恵を受け始めています。しかしこれらは始まりであって、十年後、二十年後には、今の私たちにはSFとも思われる世界が広がっている可能性があります。」*

[5] 例えば、中川氏（注３参照）も、AI倫理指針の背景として、当初は、「AI は野放図に開発すると脅威になるという感覚が生まれ、AI を人間の制御下に置くための方策として AI 倫理の議論が盛んになったと考えられる」が、「その後 AGI の可能性の研究が進み、カーツワイルのポストヒューマンのアイデアの分析が進むにつれて、AGI の可能性はまだまだ先のことであり、まして超知能の実現性は強く疑われはじめた。このような背景から、AI 倫理の当初の動機の一つであった AGIや超知能による AI 脅威論は退潮し、代わって現在ないし近い将来において重要な AI 倫理の課題が語られるようになった」としている。

[6] 例えば、その１年後となるが、DeepMindのAlphaGoが、トップ棋士の一人である李世乭九段（韓国）に4勝1敗と勝ち越したのが、2016年3月である。その後、2017年5月には、当時世界トップ棋士だった中国の柯潔九段に3連勝を果たしている。

と記載している。また、同年（2015年）6月に発表された報告書（「報告書2015」）では、「原則の検討」の必要性を記載しているが、その検討にあたっては、SF作家でもあるIssac Asimovのロボット３原則を参考にするとしている。すなわち、当時議論になっていたSFにおけるロボットのような存在であるAGI（汎用人工知能）への対応などを一部視野に入れていたと言える。なお、一方、日本の人工知能学会においては、2014年12月の第一回の倫理委員会会合を開催しているが、AI研究者の立場からは、「シンギュラリティで議論されているような「真に自己を設計できる人工知能」の実現は遠い」との認識の下、正しい現状理解の必要性について議論している[7]。

また、海外では、この時期、政府機関というよりはむしろ、民間団体が積極的に活動に取り組んでいる。その中でも、2014年に創設され、2015年1月から実質的な活動を開始したThe Future of Life Institute（FLI）の動きが注目される。同団体は、当時、人工知能、特に、汎用人工知能の開発に強く警鐘を鳴らしていたイーロンマスク氏、スティーブホーキング氏などが主要な支援者となっている団体である。同団体は、2015年1月に研究優先事項に係る公開質問状を公開し、また、2017年2月に、AI原則の一種であるいわゆるアシロマ原則を発表しているが、実際に、アシロマ原則においては、長期的課題としてではあるものの、「能力に対する警戒」「再帰的に自己改善する人工知能」に関して記述をしている[8]。すなわち、2017年時点においては、AGIの問題・懸念は長期的な話として整理されつつも、引き続き問題意識として記載されているということになる。

その後、世界各国で策定されるAI原則においても、いわゆるAGIについては、長期的な議論として整理される一方で、むしろ現実に議論されつつある問題を例示的に示したうえで、それらへの対応を含む今後の基本的方向として、各AI原則が作られるという傾向になる。例えば、2018年3月の欧州（EC）の倫理委員会の報告書では、上記アシロマ原則の動きを参照しつつも、具体的に、「自動運転」「自律兵器」「自律ソフトウェア」の事例を挙げた上で、原則の策定の必要性について記述しているが、この時点で既にAGIに対する言及は特段なくなってい。また、それを受けて2019年4月にまとめられたECのAI原則では、具体的な懸念事例として、「AIによる個人の特定・追跡」「隠れたAIシステム[9]」「基本的人権を侵害する、AIによる市民のスコアリング」「自律兵器システム（LAWS）」を挙げており、関心分野が自律性の議論から人権・公平性などの分野に推移していることが見て取れる[10]。これらを踏まえ、本ワーキングペーパーでは、AGIに係る議論は原則対象外とする。

---

[7] 第１回 人工知能学会 倫理委員会　議論まとめ（2014年12月15日）
http://ai-elsi.org/archives/262

[8] 19）能力に対する警戒： コンセンサスが存在しない以上、将来の人工知能が持ちうる能力の上限について強い仮定をおくことは避けるべきである。
22）再帰的に自己改善する人工知能：再帰的に自己改善もしくは自己複製を行える人工知能システムは、進歩や増殖が急進しうるため、安全管理を厳格化すべきである。

[9] 人は、他の人間と交流しているのか、機械（AI）と交流しているのか知ることができるべきというもの。

[10] なお、同年5月に決定された、OECDのAI原則では、その性質上、具体的事例は挙げられておらず、冒頭において、民主主義・人権、プライバシーとデータ保護、デジタルセキュリティとの関係のみについて、認識事項として記載している

＜AI原則から具体的ガバナンス・制度の検討の動き＞

上述の通り、AI原則は、少なくとも現時点では、単なる非拘束な留意事項・宣言にしか過ぎず、したがって、ガバナンスという意味では非常に弱い制度であると言える。それでは、各政策当局者は、それぞれのAI原則論を作るにあたって、今後どのようにしていくことを想定していたのであろうか。

まず、日本の総務省は、2016年の報告書においては、「OECDプライバシーガイドライン、同・セキュリティガイドライン等を参考に、研究開発に関する原則・指針を国際的に参照される枠組みとして策定」することを目的とするとしている。この趣旨は、1980年に制定されたOECDのプライバシーガイドライン（OECD８原則）が、その後の世界各国の個人情報保護法制の基本原則として取り入れられることになったことを念頭に、したがってAI原則についても、タイミングはともあれ、その後の精緻化による法制化は将来的には想定していたと思われる[11]。なお、その後、総務省は、実際にOECD等に対して積極的に働きかけを行っており、その結果が、2019年のOECDのAI原則につながっているが、OECDのAI原則においては、今後の政府政策の方向として、一般論として、「TrustworthyなAIに向けてイノベーションと競争を促進するために必要な政策、法的枠組み等を整備すべき」としているのみであり、今後の各国における法制化を明示している訳ではない。

一方、そのような中、このようなAI原則を踏まえて、一部の国・地域においては、標準化・認証の仕組みや、規制制度の検討を開始するなどの動きがあり、今後、AIシステムに対してどのようなガバナンス体制を構築していくのかが注目される。

具体的には、欧州、シンガポールでは、それらの原則をベースにしてチェックリストを作り、産業界などのAI関係者に対して、それらのチェックリストの利用を試行してもらい、フィードバックを求めることを通じて、チェックリストの精緻化を進め、標準・認証の仕組みに向けた検討が進みつつある。

また、欧州では、欧州委（EC）が2020年2月に発表したAI白書において、今後横断的な規制枠組みを作るべきとの提案を行っているのに対し、米国では、連邦政府OMBが2020年1月に発表したメモランダムにおいては、可能な限り規制的措置は入れるべきではないという方向

RECOGNISING that AI has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges;

RECOGNISING that, at the same time, these transformations may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security;

[11] ただし、翌年2017年6月の「国際的な議論のためのＡＩ開発ガイドライン案」の策定にあたっては、産業界の強い関心（反対）が強くあった。例えば、以下を参照。
日経XTECH　浅川 直輝（日経コンピュータ）「AIベンチャーの雄が総務省の開発指針に反対する理由」2017年4月10日
https://xtech.nikkei.com/it/atcl/column/14/346926/040600923/
このため、「本ガイドライン案は、非規制的で非拘束的なソフトローとして国際的に共有される指針の案として作成されたものである。」とされ、今後の規制強化については全く触れていない。

が示されている。なお、日本においては、現在特段の動きはないものの、一部法学者からは、折角AI原則を作ったので、法制化を検討すべきとの意見もある[12]。

今後のAI技術の更なるイノベーションの進展に伴い、具体的な社会への影響・課題が明らかになってくると考えられる。それに対し、単に個別問題の発生に応じバラバラに対応するのではなく、今後のイノベーションの進展に伴い発生する可能性のあるリスク・課題を見据えた上で、予め、どのような枠組みでガバナンスすることが適当なのかについてその大きな枠組みを検討しておくことが望ましい。その際に、これらのAI原則を、単なる企業等における非拘束的、自主的なチェックリストとして利用するだけではなく、現時点における各政策当局各主体の意思表明・関心表明であるとの理解の下、AIに対するガバナンス体制を構築する必要があるとした場合、それぞれの項目に対して、何故対応しなければいけないのかについて、体系立てて説明するような枠組みを検討することが必要である。

（３）「AI原則」の体系化の必要性

＜技術と倫理・社会規範＞

これらのAI原則は、AI倫理とも言われる。例えば、欧州やIEEE、UNESCOのAI原則においては、そのタイトルにおいて「倫理」という文言を利用している。一般的に「倫理」とは、「人として守り行うべき道、規律」とされ、人類が、社会・共同体としての長年にわたる持続的な発展・進化の過程において、自然発生的に生み出され、共有されてきた共通のルール・規範とであると考えられる。その内容は、高次（メタレベル）の規律から、個々具体的に取り組むべき規律まで含まれ、その規律の一部は、法制度等として明文化され、規律されることになる。この共通のルール・規範は、グローバル化の進展の中で、世界・人類共通のものも存在する一方で、各共同体における歴史的な背景を下に、共同体・コミュニティによって異なるものも少なくない。

一般的に「技術と倫理」といった場合、「技術者倫理」の分野が技術者教育の一環として確立されている。この技術者倫理の内容は、その「工学」技術の性質上、概ね、技術者の社会的責任として、各種法令の遵守（コンプライアンス）や、特に事故・安全問題や環境問題に係る未然防止に係る取組などが規律の対象になっている[13]。

一方、「技術と倫理」として、特に個別に話題になる技術分野としては、生命科学・医療技術分野、原子力・核兵器関連分野などがあげられる。これらの分野においていわゆる倫理問題が問題になるのは、おそらく、いわゆる事故・安全問題（生命・身体・財産の保全）という視点だけでなく、それ以外の、人類が遵守すべきという高次（メタレベル）の規範（以下、本ワーキングペーパーでは「社会規範」と定義する。）との関係において問題が生じる

---

[12] 例えば、新保氏（注4参照）は、「「「AI 原則は機能するか」という観点から考えてみると、これまでは原則策定の取り組みを試行錯誤し、AI を利用するにあたっての検討・研究・開発にあたり最低限必要な原則はなにかを検討する機会が多かった。今後、AI 原則ブームが到来し形式的な原則が広く普及するとなると、今後の検討事項としては、非拘束的なガイドラインや原則としての位置づけを明確にしつつ、次のステップとして、「法定公表事項」や「法定事項」として当該原則を組み込んだルール作りを考える時期に来ているのではないだろうか。つまり、実用・実装段階に向けた法規範として AI 原則の活用を考えるべきではないだろうか。」としている。

[13] 例えば、京都大学 名誉教授 嘉門雅史「技術者倫理の向上に向けて」2016年11月29日 （学術フォーラム 第8回学術シンポジウム資料）https://www.jsps.go.jp/j-kousei/data/2016_3.pdf

ためであり、そのため、これまでにこれらの技術分野において特有の各種規制・ガバナンス体制が「倫理」として講じられている。

例えば、生命科学の分野では、クローンの作成やヒト胚細胞、ゲノム・遺伝子の改変などの技術の利用に関し、「人間の尊厳」その他の人類の高次のレベルでの「倫理」（「社会規範」）の観点から、その規律の必要性について議論がなされ、それを踏まえて各技術の利用等に関する規制・ガバナンス体制としての法制度等が整備されてきている。また、医療技術分野においても、代理出産、臓器移植などの技術の利用が、同じく「人間の尊厳」などの観点から、適切か否かという議論がなされる。さらに、原子力・核兵器の分野では、一般的には平和利用は必要との認識がある一方で、人類の平和・生命の維持という高次の倫理・社会規範の観点から、重大事故や核戦争の防止などの悪用防止も含めて、原子力・核技術に対する各種のガバナンス規制が構築されてきている。

すなわち、通常の技術一般に関しても、いわゆる「技術者倫理」は課題になるものの、特に「事故・安全・環境問題」以外の高次レベルでの「倫理」規範に係るような問題が生じる場合には、「技術と倫理」に係る問題として議論がなされるようである。

＜AI技術と倫理・社会規範＞

それでは、何故、AI技術に関し、「倫理」が問題になっているのであろうか。それは、AI技術が、破壊的技術であり、単に事故・安全問題（人の生命・身体・財産の保護）だけでなく、生命科学・医療技術などと同様、それ以外に人類が共有する社会的規範、すなわち高次元での人類が有する価値観・規範に抵触する可能性があるためと考えられる。

もともと、AI技術が話題になる以前においても、デジタル技術の進展は、人類の有する社会規範との関係で抵触し、それを踏まえて、近年、多くの法令・指針などが整備されてきている。具体的には、プライバシー、セキュリティ、デジタル著作権、有害情報の扱いなどが問題になってきており、特にプライバシーに関しては、デジタル化の進展を踏まえつつ、近年、基本的人権の一つとして位置づけられつつあり、そのような認識の下で、世界各国において法制化が進んできたという経緯がある。

今回、AI技術の進展に伴い、上述のプライバシー、セキュリティなどの問題がさらに複雑化・深化するとともに、生命科学・医療技術などの「人間の尊厳」などではなく、次章において記載する通り、むしろ基本的人権、公平性・非差別性、民主主義などの人類が有する価値観・規範との関係において問題となりつつあり、そのため、「AI倫理」の必要性の認識が広がったものと考えられる。

このような中、世界各国において策定されたこれらのAI原則（AI倫理）には、非拘束な制度として、AI開発者等として「すべき事項」が含まれている。しかしながら、その中には、人類が共有して取り組むべき事項としての高次レベルの社会規範としての事項と、そのような規範を守るためにAI技術に対するガバナンスとして取組むべき事項が、両方混在して含まれているのが現状である。

今後、規制の必要性の有無を含めガバナンスの在り方を議論していくためには、社会規範（高次元の価値基準）として取り組むべき事項と、ガバナンスの一環として取り組むべき事項に分けて整理することが必要である。このような認識の下、今後のAIに係るガバナンスの方向の検討に資するべく、次章以下においては、これまで世界各国で策定されているAI原則を体系化し、分析する。

３．AI原則の分析・体系化（試案）

（１） 分析の枠組みーイノベーション、社会制度（ガバナンス）、社会規範

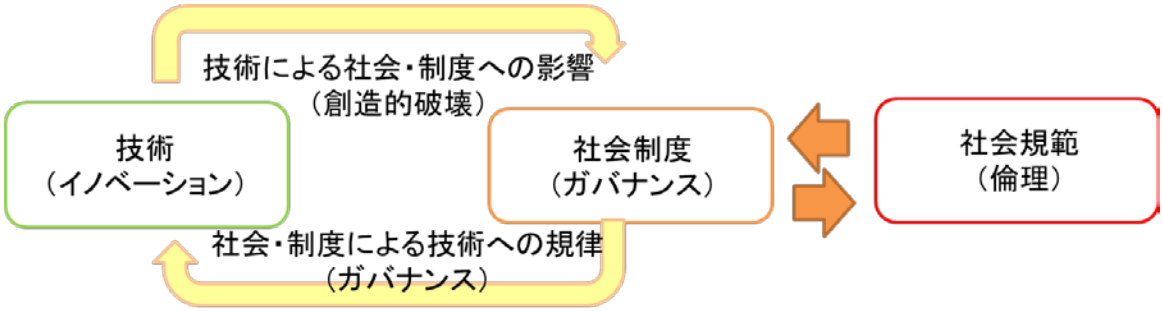
本章においては、前章における問題意識の下、まずは、技術・イノベーションに対する社会規範とガバナンスの位置づけを整理する。

一般的に、新たな技術が創出され、イノベーションが進展すると、その程度にもよるが、何らかの社会制度を創出することにより、当該技術を律しようとする動きが生じる。このような社会制度創出の動きは、イノベーションの本質でもある。今回の議論するAI原則の作成・提示は、ある意味での制度の創出の一端であると位置づけられる。

そもそもイノベーションとは、シュンペーターの定義の通り、「創造的破壊」である。一般的に、シュンペーターのいう創造的破壊は、経済発展の視点から語られる。すなわち、例えば、新たな効率的な方法が生み出され、その普及が進むと、それと同時に古い非効率的な方法は駆逐・破壊され、その一連の新陳代謝のプロセスを通じて、経済発展が進展するというメカニズムである。

しかしながら、その技術・イノベーションの「破壊性」によっては、経済だけでなく、それを支えている社会全般あるいはその社会の有する価値観に対しても大きな影響を及ぼしうる。その際、技術が社会に受容されるには、人類やその共同体が共有する社会規範（倫理）に照らして、それに適合する形で、既存の制度の見直しや新たな制度構築など、当該技術に対する適切なガバナンス体制が模索されることになる。一方で、このような技術に対する制度・ガバナンス体制は、その後の技術・イノベーションの進展・方向に対して大きな影響を与える。そのような意味で、「技術・イノベーション」と「制度・ガバナンス体制」の共進化が行われる。図表３に示す。

（図表３）技術・イノベーションと制度・ガバナンス、社会規範との関係（枠組み）



実際に、本枠組みに照らして、過去の破壊的な技術・イノベーションの事例を考えると、いずれも、社会・制度に対して大きな影響を与え、新たな制度・ガバナンスが構築されることにより、技術に対する規律がなされ、技術と制度の共進化が進んできた。例えば、20世紀初頭に普及が開始された自動車の場合は、当初は安全の問題が発生し、それに対して各種の交通ルール、損害補償制度等が整備されるとともに、それに伴い自動車安全に係る技術も進

展した。その後、環境の問題が発生し、その対応として各種の環境規制が導入されるとともに、環境配慮型の自動車の開発が進んだ。いずれも、社会における安全や環境保全などといった社会規範の明確化というプロセスと並行して、技術・イノベーションと社会制度・ガバナンスが共進化してきたものと位置づけられる。

なお、その際、高いメタレベルの倫理である社会規範自体は、必ずしも世界共通であるとは限らない。もともと、倫理・社会規範とは、人類が、それぞれの社会・共同体としての長年にわたる持続的な発展・進化の過程において、文化の一部として自然発生的かつ進化的に生み出され、共有されてきた共通のルール・規範とであると考えられる。このため、グローバル化が進展したとは言え、例えば、欧州、中東、東アジアなども含め、各共同体によって異なるものであり、また、国の体制によっても異なるであろう。

また、このようなメタレベルの社会規範は、時代によって変化するものであり、現在でも変化しているものと考えられる。例えば、現在、包摂的成長、持続的発展というキーワードが流行りであるが、これらは最近において概念化されたものである。また、人権という概念でさえ、フランス革命以降の概念であり、男女平等（例えば参政権など）も19世紀後半の概念であり、また、米国で人種差別が撤廃されたのも20世紀半ばである。さらに、このような社会規範の概念は、必ずしもトップダウンで決定されるものではなく、共同体内における様々な議論の中で、ボトムアップにより生み出され、社会通念として次第に共有・合意されていくものと考えられる。

なお、このような中、技術・イノベーションの進展自体が、この「社会規範」（倫理）に対しても大きな影響を与えるものと考えられるが、そのメカニズムについては、将来の研究課題であるといえる。例えば、プライバシーという概念は、デジタル技術の進展に伴って概念化が進展し、その保護は、近年、基本的人権の一つとして位置づけられるようになってきているとともに、その内容は、現時点でも、技術とともに進化中であるとも言える。

（２）世界の各AI原則の分析・整理

上記枠組みを踏まえて、以下において、世界各国のAI原則を分析する。具体的には、AI原則には、基本的には、AIの開発者・利用者等が遵守すべき事項が記載されているが、目指すべき社会の在り方としての「社会規範」として遵守すべき事項と、それを実現すべき「ガバナンス」として遵守すべき事項が混在しているとの認識の下で、以下の手順にて、それを整理・体系化する。

- まずは、世界各国のAI原則論を、社会規範・ガバナンスを意識しつつ、項目別に整理することにより、全体像を把握する。
- それぞれの項目の内容を読み込み、記載されている各留意事項まで把握した上で、それらの事項ごとに、社会規範、ガバナンス（組織、システム）のいずれに該当するかを分類する。
- その上で、その内容を踏まえて、その事項が該当する項目を、必要に応じ移動を行うことにより、各項目の並び替えを行い、全体を整理・体系化する。

＜世界のAI原則の全体像＞

まずは、図表１のうち、世界各国・機関（8カ国・機関）のAI原則について、まずはその各AI原則の項目を整理する。

具体的には、世界各国の意見を最も多く集約したと考えられるOECDのAI原則（5原則）の項目をベースに、「社会発展」「人権・公平性」「透明性」「安全・信頼性」「説明責任」の５項目を立て、それ加え「環境整備」の項目を加える。なお、このうち、「社会発展」「人権・公平性」は、社会として目指すべき社会規範的要素が強いのに対し、「説明責任」「安全・信頼性」「透明性」はそれを実現するためのガバナンス的要素が強く、特に「説明責任」はすべてのガバナンスを統括する概念であると言える。また、各国のAI原則は、概ね、AI開発者・利用者が遵守すべき事項が記載されているが、AI原則によっては、他の関係者との連携も社会全体での取組みの視点を含むものあり、そのような社会全体として取り組むべき事項に係る項目を「環境整備」とした。

上記項目に基づき、OECD以外の７か国・機関のAI原則の各項目を、それぞれ最も近い内容と思われるものをあてはめて、整理を行ったものが図表４である。同じ行にある各AI原則の項目の内容が他のAI原則の項目の内容と全く同じという訳ではないが、概ねどのAI原則も同じような内容を項目に挙げていることがみてとれる。

なお、OECDのAI原則においては、「プライバシー」は「2.人間中心の価値と公平性」に含まれるが、他の全てのAI原則に記載があるので別の行に整理した。また、「人間の代理・監督、制御可能性、自立性尊重」といった項目も、別の行に整理した。

（図表４）世界各国・機関（8カ国・機関）のAI原則の全体像

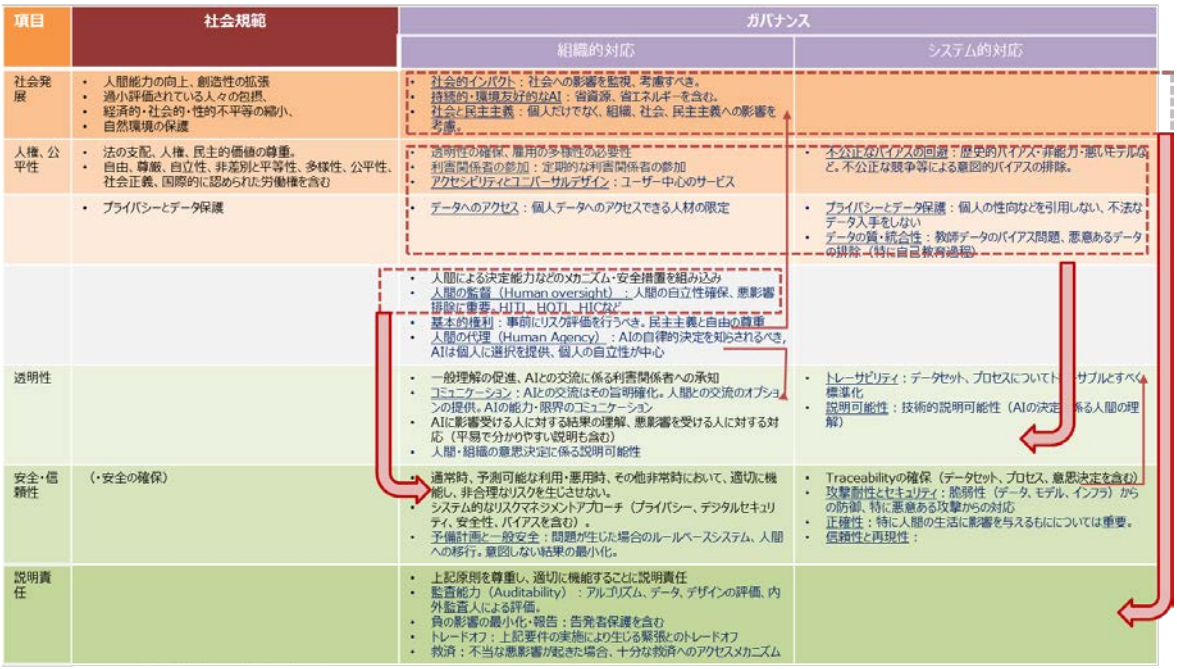
項目	OECD	EU	日本（CAO）	日本（MIC）	カナダ（U-M）	IEEE	中国	UNESCO
社会発展	1 包摂的成長、持続的成長、幸福	6.社会的・環境的幸福	1.人間中心		1.幸福 10.持続的発展	2.幸福	1.調和・友好 3. 包摂・共有	7.人類の繁栄のために 8.均整 10.持続可能性
人権、公平性	2.人間中心の価値と公平性	5.多様性・非差別・公平性	6.公平性	7.倫理	5.民主的参加 6.平等性 7.多様性包摂性	1.人権	2. 公平・公正	11.多様性と包摂 15.公平性
		3.プライバシー・データガバナンス	3.プライバシー確保	6.プライバシー	3.プライバシー、親密性	3.データエージェンシー	4.プライバシーの尊重	12.プライバシー
		1.人間の代理・監督		3.制御可能性	2.自律性尊重	7.誤用の気づき		9.人間の監督と決定
透明性	3.透明性と説明可能性	4.透明性		2.透明性		5.透明性		16.透明性と説明可能性
安全・信頼性	4.頑健性、セキュリティ、安全性	2.技術的頑健性、安全性	4.セキュリティ確保	4.安全 5.セキュリティ		4.効率性 8.能力	5.セキュリティ・制御可能性	17.安全性とセキュリティ
説明責任	5.説明責任	7.説明責任	6.説明責任、透明性	9.アカウンタビリティ	8.慎重性 9.責任	6.説明責任	8.アジャイルガバナンス 6.責任の分担	14.マルチステークホルダーと適合的ガバナンス 18.責任と説明責任
環境整備			2.教育リテラシー 5.公正競争確保 7.イノベーション	1. 連携 8.利用者支援	4.連帯		7.開放・協力	13. 理解とリテラシー

＜個別項目の内容の整理＞

次に、それぞれの項目の内容につき、特にOECDのAI原則とEU（EC）のAI原則に記載された詳細事項に立ち入って分析・整理する。なお、OECDのAI原則は、比較的高いレベルから記載している内容が多いのに対し、ECのAI原則は、実際には「要件」であり、ガバナンス対応としての具体的な記載内容が多い。このため、全体像を最低限把握するため、この２つのAI原則を分析対象とする。したがって、これらの記載内容は、世界全体で議論されている内容を必ずしも網羅したものではないこと、また、一方で、EUの原則が中心となっているため、世界で合意された内容でもないことに留意することが必要である。

具体的には、この2つのAI原則に記載されている詳細事項について、筆者の判断により、社会規範に係る記述とガバナンスに係る記述、また、ガバナンスに係る記述に関しては、さらに、組織に係るものとシステムに係るものに分類すると、図表５となる。

（図表５）OECD・EUのAI原則の詳細記載事項の分類（試案）[14]



この図表５を踏まえて、「社会発展」「人権・公平性」の項目は、社会規範にかかるものと、また、「透明性」「安全・信頼性」「説明責任」に係る項目は、ガバナンスに係るものとして、改めて整理をするものとし、具体的には、以下の通り整理を行う。

【社会規範について】：社会規範に係る項目として、「社会発展」「人権・公平性」（プライバシーを含む）の項目にある各事項を整理し直すとともに、新たに「安全性」を項目として追加する。

- 「社会発展」を社会規範の一つとして整理する一方、そのガバナンスの部分に記載のある事項（社会的インパクトの評価など）やECの「人間代替・監督」に記載のある「基本的権利に係るリスク評価」に係る事項等は、「説明責任」の一部として位置づける。
- 「人権・公平性」も社会規範の一つとして整理する一方、そのガバナンス部分に記載のある事項は、広い意味での「透明性」に係る内容及びデータガバナンスに係る内容であることから、まとめて、「透明性・説明可能性」の項目の事項として位置づける。
- 一方、「安全性」に係る項目を追加し、「安全・信頼性」に係る取組に対応するような社会規範として位置づける。なお、生命、身体、財産の保護のなどに係る「安全性」は、本来、人類の基本的権利であり重要な社会規範であるし、また、世界の

[14] 注：黒字部分は、OECDのAI原則。青字部分は、EUのAI原則。赤字の矢印部分が、筆者の判断により項目間の移動を行った事項。

他のAI原則の一部にはキーワードとして「安全」に係る記述はあるが、OECD、ECのAI原則には、明示的な記載はない。おそらく、「安全」は当然のことであるとみなし、「安全・信頼性」に係る各事項を記載すれば十分意図が伝わるものと考えたものと推測される。

【ガバナンスについて】：「透明性」「安全・信頼性」「説明責任」に係る項目は、主に「ガバナンス」に係る項目として整理し直す。

- 「透明性」については、上述の通り、「人権・公平性」のガバナンス部分に記載のある項目も加えて、「透明性・説明可能性」とし、主に「人権・公平性」に資するガバナンス上の取組として整理する。なお、ECの「人間代替・監督」に記載のある「人間の代理」については、OECDの「透明性」に同等の記載のあることから、本項目の事項として整理する。また、OECDの「頑健性、セキュリティ、安全性」に記載のある「トレーサビリティの確保」は、ECの「透明性」に同等の記載があることから、同様に本項目の事項として整理する。
- 「安全・信頼性」については、「信頼性・制御可能性」とした上で、上述の通り、主に「安全性」という社会規範に資するガバナンス上の取組として整理する。なお、OECD、ECの「人間の監督」については、「安全・信頼性」として整理する。
- 「説明責任」については、OECDのAI原則に「「上記（１～４）の原則を踏まえて、アカウンタブルでなければいけない」と記述されている通り、ガバナンス全体に係る概念として整理する。

＜「人間の監督」に対する考え方＞

なお、ここで、ECのAI原則のトップに記載のある「人間の代理・監督」の項目の取り扱いについて考察する。

ECのAI原則では、「AIは、人間の自立性をサポートするものであるべき」（人間の自立性の尊重）との考え方（上位概念）の下で、「人間の代理・監督」として、①基本的人権（リスク評価の必要性など）、②人間の代理（Human Agency：AIの決定であることを人間に知らせるべきなど）、③人間の監督（Human Oversight：AIに対する人間の監督による人間の自立性確保、悪影響排除など）を記載している。また、OECDのAI原則においても、「透明性・説明可能性」において②の同旨が、また、「人間中心の価値・公平性」において③の同旨が記載されている。

この「人間の代理・監督」という概念、特に③「人間の監督」については、二つの見方がある。一つは、「AIは、人間を超えて意思決定してはならない」という社会規範を守るため人間が監督しなければならないという見方であり、もう一つは、当面AIは人間を超えることはできない（AIには限界がある）ので、安全性の確保等の社会規範確保のためには、最終的に人間が監督しなければならない、という見方である。

前章で議論した通り、「社会規範」は、共同体によって異なる可能性がある。このうち、前者の見方は、言い換えれば、将来的なAGIを念頭におき、人類社会は人間自らの能力を超える存在（AI）を作り出してはいけない、人間を作った神に対する冒涜にあたるという西洋

のキリスト教史観が背景にあると思われる[15]。本ワーキングペーパーでは、将来的なAGIを念頭においておらず、AIシステムとは人間の監督の下で運用される情報システムであるとの前提で議論することも踏まえ、後者の立場を取り、安全性の確保等のために、システムには制御可能性を確保すべきとの視点で整理する。

（３）AI原則の構造化・体系化（整理結果）

上記の整理を踏まえて、AI原則を、概念的に「社会規範」と「ガバナンス」に分けて整理したものが、図表６、図表７になる。これらを説明すると以下の通り。

【ガバナンス全体】

- 社会発展を中心とし、それに加えて、安全性、人権・公平性等も配慮した社会規範を実現するために、説明責任を中心とするガバナンス体制を構築する。
- 説明責任には、リスク・インパクト評価、監査、報告、利害関係者参加などが重要となるが、信頼性・制御可能性、透明性・説明可能性についても、説明責任の一部に含まれる。

【安全性のためのガバナンス】

- 安全性の確保のため、信頼性・制御可能性を中心とするガバナンス体制を構築する。
- その際、責任体制、人間による監督、予備計画などの策定に加え、システム・技術の観点からは、信頼性・品質・正確性、頑健性・セキュリティ・安全性、誤用・悪用防止・制御可能性などが重要である。

【人権・公平性のためのガバナンス】

- 人権・公平性の確保（プライバシー保護を含む）のため、透明性・説明可能性を中心とするガバナンス体制を構築する。

[15] 例えば、オラフ・グロス 、 マーク・ニッツバーグの「新たなＡＩ大国　その中心に「人」はいるのか？」講談社（2019/12/6)には、日本の人工知能研究者に対するインタビューの結果として、以下の内容を記載している。

「日本の産業技術総合研究所（AIST）で人工知能研究センター長を務める辻井潤一は言う。「海外ではモンスターのようなイメージを持たれがちですが、日本では、ロボットは人間の護り手か友達のようなものなんです」と。」（略）「日本の社会が先進的テクノロジーを受け入れるのには、人口動態やポップカルチャーよりもさらに根深い理由がある、と。それは東洋では昔から当たり前の基本的な哲学によるものだ。西洋思想と違って、東洋では人間が特別でない。」（略）「「私たちには、一神教の神のような「創造主」という概念がありません」と辻井は言う。「西洋文明は常に『人間は神のコピーであり、人間には特権が与えられている』と考えています。アジアの文化には、そうした考えがありません。動物から人間まで、緩やかにつながっているのです。」と。」（略）

「東京大学・知能情報システム研究室の國吉康夫教授は、「人間そっくりのヒューマノイドは、人工知能の今後の性向のために重要だ」と強調している。」（略）「「私たちは、人間のようなものを作ろうとしています」と教授は言う。「それが悪いことだとか、恐ろしいことだとは思いません。そこがおそらく欧米人と日本人の違いなのでしょう。欧米人の多くは、人間に匹敵するような別個の存在を要因できないのです」と。」

- その際、透明性の確保、周知、関係者との対話、意思決定の多様性に加え、説明可能性・トレーサビリティ、データの質確保、データガバナンスなどが重要である。

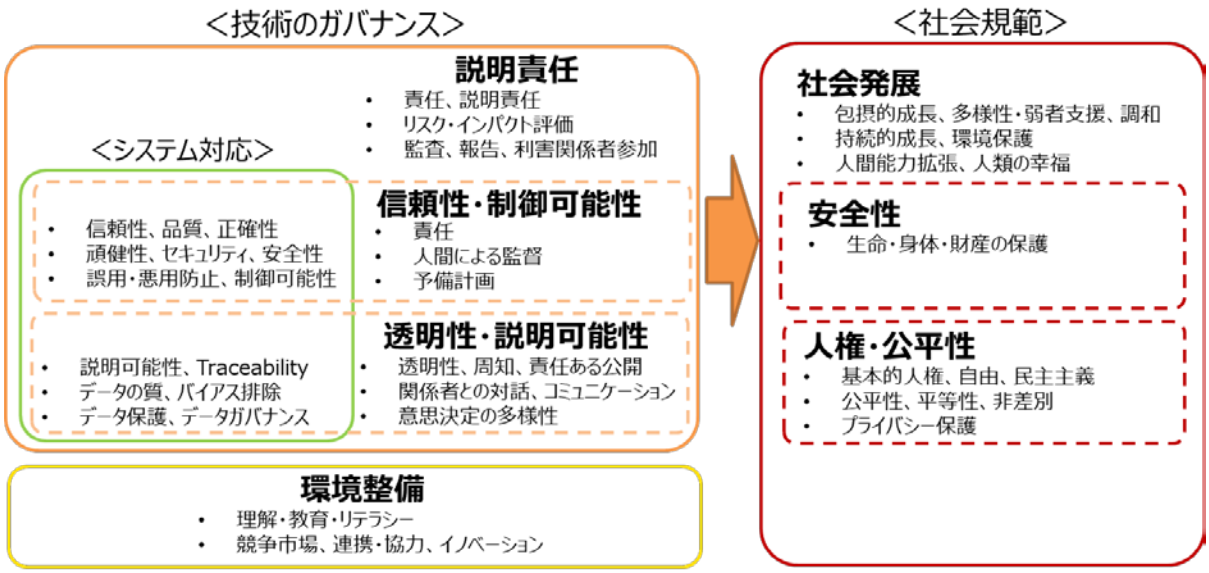
【環境整備】

- それに加えて、理解・教育・リテラシー、競争市場の確保と連携・協力、イノベーションなどの環境整備が重要である。

（図表６）AI原則の構造化試案（社会規範とガバナンス）

	全般的	主に安全に係るもの	主に人権・公平性に係るもの
社会規範	• 包括的成長、弱者支援、調和 • 持続的成長、環境調和 • 人間能力拡張、幸福	• 生命・身体・財産の保護	• 基本的人権、自由、民主主義 • 公平性、平等性、非差別 • プライバシー保護
ガバナンス	全般的項目	主に信頼性に係るもの	主に透明性・データに係るもの
組織対応	• 説明責任 • リスク・インパクト評価 • 監査、報告、利害関係者参加	• 責任 • 人間の監督 • 予備計画	• 透明性、利用周知、責任ある公開 • 関係者への説明・対話 • 意思決定での多様性
システム対応		• 信頼性、品質、正確性 • 頑健性、セキュリティ • 誤用防止、制御可能性	• 説明可能性、トレーサビリティ • データの質・バイアス排除 • データ保護、データガバナンス
環境整備	• 競争市場、連携・協力、イノベーション		• 理解、教育、リテラシー

（図表７）AI原則の構造化試案図（社会規範とガバナンス）



なお、上記整理は大きな枠組みのみを示すものであり、今後さらに精緻化していくことが望まれる。特に、個々の取り組むべきとされる事項については、OECD及びECのAI原則に記載のあるもののうち、主なもののみを記載したものであり、他のAI原則の記載事項などを踏まえると、さらに追加・整理することが可能である。

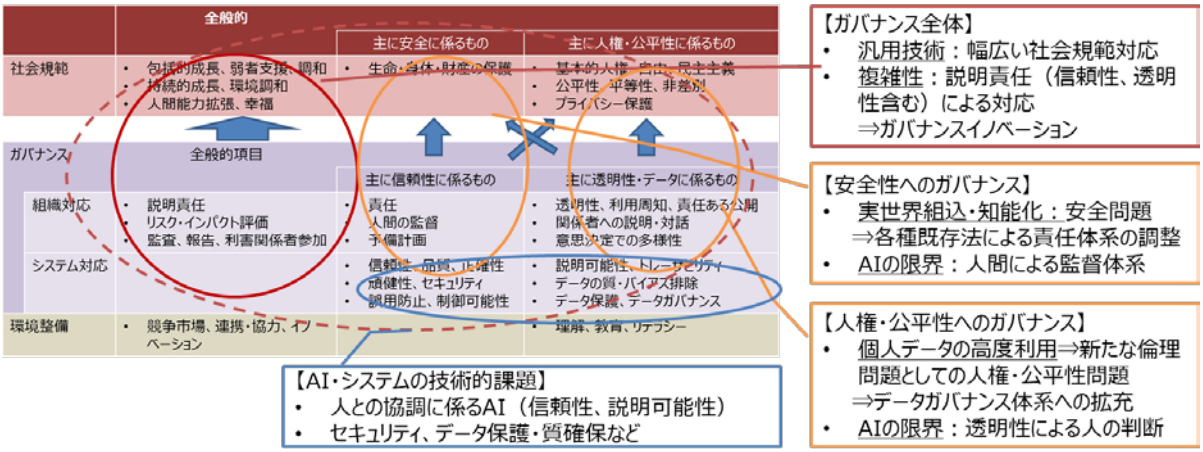
４．AIガバナンスの今後の方向（考察）

（１） AIガバナンスに係る全体枠組み

本章では、前章までのAI原則の体系化・分析を踏まえて、デジタル・AIシステムに対するガバナンス体制の在り方について考察する。具体的には、図表８に示す通り、以下の４つに分けて、考察を行う。

- ガバナンス体制全体の枠組み（（１）節）
- 安全性に対するガバナンス（（２）節前半）
- 人権・公平性に対するガバナンス（３節後半）
- ガバナンス対応に必要なシステム・技術課題（（３）節）

（図表８）今後のAIガバナンスに係る各項目



＜AI原則にみるAI全体に係る社会規範とガバナンス＞

AI・デジタル技術は、汎用性の非常に高い技術であり、様々な分野に応用が可能である。このため、持続的発展、包摂的発展、人間の拡張など、様々な人類の福祉の向上を目的にも活用できるのが特徴である。この点は、従来、倫理問題が大きく議論されている生命技術、原子力技術とは様相が異なる。

一方で、これらの技術に係る副作用として、右側にある安全性や人権・公平性などの各種の問題が生じうる。さらには、悪意ある利用がなされる可能性もある。そもそも、技術がその副作用として各種問題を生じうることは他の技術においても同様であるが、このようなAIシステムに対しても、他の技術と同様に、その利用分野等に応じて、何らかの適切なガバナンス体制の構築が求められる。

上述のAI原則の体系化を踏まえると、全般的なガバナンス体制としては、AIシステムの開発・利用者等に対して、説明責任（アカウンタビリティ）を課すことが基本となる。具体的には、各国・機関のAI原則に記載されている内容・事項としては、リスク・インパクト評価の実施、監査、報告、利害関係者の参加などがあげられている。またそれに加えて、その内容によっては、右側に記載のある信頼性・制御可能性、透明性・説明可能性に係る対応も、そのガバナンスの一環として行うことが求められる。

＜具体的なガバナンスの方向の検討～ガバナンスイノベーション＞

それでは、このようなAI原則に係る枠組みを念頭に置きつつ、政府は、具体的にどのような規制を含むガバナンス体系を構築していく必要があるのであろうか。全体の技術・イノベーションと制度・ガバナンスの共進化に係る枠組み（図表３）を踏まえると、イノベーションを促進するような規制・ガバナンス体制が求められ、具体的には、その要件としては、以下の２点が重要になると考えられる。

①リスクに応じた柔軟性の高い規制・ガバナンス体系

- まずは、当該システムによる影響度合・リスクに応じて、必要な規制・ガバナンス体制を構築していくことが適切と考えられる。また、AI技術・イノベーションは、今後も引き続き変化し続けることを想定すると、技術の進展に応じたフレキシブルな対応が可能なガバナンス体制が適切であると考えられる。
- なお、規制に係る考え方の一種としては、「予防原則」に基づく規制の議論がある。「予防原則」とは、化学物質や遺伝子組換えなど環境に重大かつ不可逆的な影響を及ぼす恐れがある場合、科学的に因果関係が十分証明されない状況でも、規制措置を可能にする制度や考え方のことである。このような規制は、AGIあるいは「再帰的に自己改善する人工知能」への対応としてはありうるが、一般的なAIシステムに適用すべきものではないと考えられる。

② 技術要件型規制から目標設定型ガバナンス体系へ（ガバナンスイノベーション）

- 一般的には、通常の規制においては、その規制の対象とされる利用分野において、対象とされるシステムの技術的な要件を国・政府が定め、その遵守を国・政府が確認するという手法が主流である。しかしながら、この技術要件型の規制の場合、要件の規定方法にもよるが、同じ目的を達成するにあたっても、技術のイノベーション、創意工夫を妨げることになる可能性がある。さらに、AI・デジタル技術は、非常に複雑化しており、国なり公的部門が、規制対象者となる企業等の個々のシステムの内容を把握することは困難であり、個々の技術規制を設けること自体も困難になっている（企業等でさえ、自らのシステムの挙動を全て把握できないこともある）。このため、規制を行うにあたっては、単純に技術要件規制を行うのではなく、目標設定を行った上で、上述のような説明責任、マルチステークホルダーの関与、透明性の確保を課すといったガバナンスの方法の導入を考慮する必要がある。
- もちろん、個別の特定可能なサブモジュールについては、その必要性は検討する必要はあるものの、規制を行うという選択はありうる。実際に、Deep Fake[16]や、顔認

[16] 例えば、米国カリフォルニア州では、2019年10月3日、州知事は、政治候補者への信頼を傷つけたり有権者を欺いたりすることを狙った配布を違法とする議会法案、及びの、本人の同意なくポルノコンテンツに使われた場合、そのディープフェイクを作成した人物を提訴する権利を州民に認める法案に署名したと報道されている。

CNET Japan「カリフォルニア州、政治やポルノのディープフェイクを法律で規制へ」2019年10月7日
https://japan.cnet.com/article/35143600/

また、中国では、中国サイバースペース管理局（CAC）は、2019年11月29日、人工知能（AI）や仮想現実（AR）を利用して作成されたコンテンツには、その旨を明記することを義務付け、明記がない

証技術の利用[17]に関しては、一部規制が導入されつつある。ただし、これらにおいても、技術の発展を阻害しないような適切な規制が求められる。

AIシステムに対する規制に関しては、2020年2月にECが発表した「AI白書」では、新法も含む規制強化の方向を打ち出しているのに対し、2020年1月に米国ホワイトハウスOMBが発表した「AI応用の規制のためのガイダンス（ドラフト）」では、非規制的措置への指向を強く打ち出すなど、欧米で全く異なる対応方針を示している。

しかしながら、それぞれにつき、より詳細を見ると、その程度論はともかくとも、方向性としては、全く異なる訳ではなく、両者が一致する面も少なくないと考えられる。

- 欧米ともリスクに応じた規制を志向している。欧州のAI白書では、「基本的人権へのリスク（プライバシー、非差別など）」と「安全性・責任関係のリスク」に分けるとともに、リスクの高い分野については法的枠組みの対象とする一方、リスクの低い分野では、自発的ラベリングなどを求めている、米国のガイダンスにおいても、非規制的措置を指向するとしつつも、リスクアセスメント、柔軟性、その他の10の項目を検討するとしており、規制的措置について完全に排除している訳ではない。
- また、規制する場合において、欧米とも単純な技術要件型の規制を指向している訳ではない。欧州は、要件として、データ、記録、透明性などを重視し、執行においても事前適合性評価を指向している。米国では（必ずしも明記されていないものの）科学的な統合性と質を重視するとともに、今後の適合性評価手法の開発について記載している。

なお、日本では、特に後段（②）のガバナンス体制に関連して、経産省は、2020年7月、報告書「GOVERNANCE INNOVATION　Society5.0の実現に向けた法とアーキテクチャのリ・デザイン」[18]を発表している。本報告書では、特にデジタル分野における規制について、「国がルールの設計から監督と執行までを一手に担う従来型のモデルから脱却し、企業がルール設計とモニタリング、エンフォースメントの中心的な担い手となっていく」として、民間企

場合は刑事罰の対象とするとの新たな規制を導入する（施行は、2020年1月1日）と発表したと報道している。

ロイター「中国、偽ニュース取り締まり強化へ「ディープフェイク」に警戒」2019年11月30日
https://jp.reuters.com/article/china-fakenews-cac-idJPKBN1Y326W

[17] 例えば、サンフランシスコ市では、警察など市の53機関での顔認識技術の利用や顔認識技術で取得された情報の利用を禁止する条例が、2019年6月30日から施行。

なお、欧州では、EDPD（欧州データ保護会議）が、GDPRに基づき、2019年7月10日に「ビデオ機器を通じた個人データ処理に関するガイドラン」案を公表している。
国際社会経済研究所主任研究員小林雄介「欧米におけるカメラ・顔認識サービスと規制動向」2019年10月29日
https://www.i-ise.com/jp/information/report/2019/20191029.html

[18] 経済産業省Society5.0における新たなガバナンスモデル検討会「GOVERNANCE INNOVATION　Society5.0の実現に向けた法とアーキテクチャのリ・デザイン」2020年7月13日
https://www.meti.go.jp/press/2020/07/20200713001/20200713001.html

同報告書では、イノベーションを促進するガバナンス：Governance for Innovation）。イノベーションに対するガバナンス：Governance of Innovation）、イノベーションを活用したガバナンス：Governance by Innovationの観点から、新たなガバナンスについて議論を行っている。

業による自主的な取組や、多様化した価値観を有するコミュニティや個人による積極的な関与を重視した、マルチステークホルダーによるガバナンスモデルとして、概ね上述（②）に沿った方向の体制の議論を行っている。今後、AIに対するガバナンス体制の構築にあたっては、単に規制するか否かではなく、このようなガバナンスイノベーションの視点を踏まえ、どのようなガバナンス体系を選択・導入していくのかについての検討が求められる。

（２）安全性と人権・公平性に係るAIガバナンスの方向

次に、上記の全般論に加えて、AIシステムのガバナンスにおいては、大きく分けると、安全性に対するガバナンスと、人権・公平性に対するガバナンスが、重要になる。特に、後者については、デジタル・AI技術に特有で近年新たに浮上してきている問題であると言える。なお、実際に、欧州委員会（EC）も、この二つのリスクは分けて議論すべきとしている。

- 安全性（生命、身体、財産の保護）については、そのリスクの大きなAIシステムを対象とした分野特有のガバナンスが求められる。特に、既存の各分野の安全に係る法規制体系を中心にした見直しが求められる。
- 人権・公平性については、個人データを扱うようなAIシステムを対象としたデータ特有のガバナンスの在り方への対応が求められる。特に、現在のプライバシーを中心としたデータガバナンス体制について、その対象の拡充を含めた対応の必要性の検討が求められる。

以下においては、AI・デジタル技術の特徴も踏まえた上で、それぞれの社会規範の確保に向けたガバナンスの在り方について、考察する。

①安全性に係るガバナンス：信頼性・制御可能性など

＜既存の法体系とガバナンス体制の調整・見直し＞

生命、身体、財産の保護といった安全性の問題は、人類が長らく醸成してきた社会規範であり、この実現のために多くの技術がガバナンスの対象とされてきている。前章に記載の通り、世界のAI原則に必ずしも全て明示されている訳ではないが、今後、AIシステムがより知的になり、現実世界に組み込まれ、また、社会のインフラ基盤としての役割を担うようになると、この安全性という社会規範は、「技術と倫理」との関係からも、現実世界において重要な課題となりなる。

このような安全性に対するガバナンスとしては、これまで、一般法（民法・刑法）に加え、特にリスクの大きな分野を中心に各種既存法が整備されている。近年のデジタル化の進展に伴い、既にシステムの信頼性、サイバーセキュリティの問題が課題となってきているが、これに加えてAI技術の進展に伴い、さらにシステムが複雑化すると予想される。このため、上述（１）に記載したような新たなガバナンス体制を既存の法体系に組み込むべく、見直し・調整を進めることが必要になることが考えられる。

具体的には、まず、現状の法体系においても、AIシステムに限らず、何らかのシステムの不具合等により、事故等が生じた場合には、民法、刑法の一般原則に基づいて、責任関係が問われることになる。すなわち、一般的にシステムの不具合で事故等が生じた場合は、原則

として、そのサービス提供者に過失があるかということで責任の有無が判断され、その際、過失とは予測可能性があったかが鍵になる。このような基本的考え方は、原則として、AIシステムにおいても踏襲・適用されるものと考えられる[19]。

一方、自動運転・医療など特に人の生命に関わるものなどリスクの高い分野においては、安全性の確保に係る技術要件規制も含めて、これまで各個別法の整備がされてきている。このような法令に関しては、AIシステムの導入にあたって、上述の（１）に示すような説明責任、利害関係者への説明、透明性の確保なども含む規制・ガバナンス体制の見直し・調整が必要になってくるものと考えられる。

＜人間の監督を含む多重の安全性と予測可能性・信頼性の向上＞

AI技術は、「知的」であり、何らかの自律性を有する。このため、AIシステムが導入されるプロセスにおいて、従来その役割を担っていた人間に代わり、AIシステム自らがある程度の意思決定・判断などの自律的行動を行うことが可能になる。このため、安全性に係る行政上の責任に係る対象者を「人」（例えば、運転者、医師など）に規定している個別法等においては、必要に応じて、これらをAIシステムの運用者に置き換えるべく見直しをすることが必要になる。

一方、（AGIを想定せずに）当面のAI技術を考えた場合、AIの意思決定能力には限界がある。AI技術は、人がルールベースでプログラミングをする場合にせよ、人が一定のデータを与えて機械学習を行い作成する場合にせよ、一定の範囲・枠組み（フレーム）の中における各種ケースを想定しつつ、人が作成する計算機械にしか過ぎない。したがって、状況に応じて臨機応変に判断が可能な人・組織と比較して、AIシステムは当該フレームの枠を超えて生じた事象に対して適切な意思決定を行うことができない[20]。このため、当該AIシステムが利用・適用されている場面やその自律性の程度にもよるが、安全性等の確保を図るためには、基本的には、前章で述べたような「人間による監督」（Human Oversight）を組み込み、当該システムに管理する組織・人間が最終的な管理責任を担うことが必要となる。その際、特にリスクの高い分野においては、単に人間・組織の管理に任せるだけではなく、システムが機能しなくなった各種の場合を想定し、予備計画の事前策定を含め、多重の安全の枠組みを構築することが求められる。

また、当該システムを管理する組織等がシステムに係るリスクに対し責任を担う前提として、当該AIシステムの挙動について正確に把握するなど予測可能性を把握し、それに応じてシステムの信頼性を確保するための取組を行うことが必要となる。特に多量のデータを用い機械学習によって得られたソフトウェア・アルゴリズムについては、人がルールベースで作成されたソフトウェアとは異なり、そのシステムの作成者でさえ、その挙動について十分に理解できない場合がある。このためには、AIシステムの品質評価を行う手法が重要になる。

さらに、そのリスク対応の一環として、頑健性（Robustness）の確保やセキュリティへの対応も必要となる。セキュリティとは、一般的に、機密性（情報漏洩等）、完全性（情報改

[19] もちろん、このような法体系は、安全性に限らず、人権・公平性に関連して特定の被害が生じた場合には同様に適用されうるものである。ただし、下記に記載する通り、人権・公平性に関しては、何を具体的な被害として特定するかが法令上明確になっていない場合が多いと考えられる。

[20] また、トロッコ問題のように、そもそもどういう対応をとることが正しいのかについての社会規範が定まっておらず、人間でさえも何が倫理的に「正しい」のか判断できない場合も存在する。

ざんなど）、可用性（障害対応など）を維持することであり、特に機密性については、プライバシーや個人情報の保護等のためにも重要となる。セキュリティの問題は、AIシステムに限った議論ではないが、教師データの改ざんなどAI技術特有のセキュリティ問題も存在するため、それらに係る対応も必要となる。

②人権・公平性に係るガバナンス：透明性・説明可能性など

＜プライバシー権から人権・公平性へのデータガバナンスの拡充＞

上述の「安全性」に係る社会規範は、長らく人類が醸成し、対応してきた課題であるのに対し、「人権・公平性」に係る社会規範は、近年のデジタル・データの時代を背景に新たに浮上してきた比較的新しい課題であり、ある意味、「AI倫理」の中心的な課題として位置づけられる最先端の部分であると言える。キーワードとしては、基本的人権の尊重、プライバシーの尊重、公平性、平等性、非差別性、民主主義などであり、狭義でのいわゆる「倫理」問題であると言える。

このうち、プライバシーの問題は、近年のデジタル化の進展に伴い、概念的には基本的人権の一つとして新たに位置づけられるとともに、個人データの扱いに係るガバナンスの在り方について議論が進展してきている。近年のビッグデータ化の進展とAI技術の発展に伴い、個人のIDに紐づいて入力される属性データが大幅に増大するだけでなく、顔認証技術等の利用により、画像情報に紐づく多様なデータと個人のIDとが非意図的に関連づけられることが可能となるなど、その対象とするガバナンスの範囲は大幅に拡大しつつある。

それに加え、個人のIDと紐づく情報の扱いは、プライバシーの問題を超えて、公平性・非差別性・平等性などの観点から問題が生じてきている。AIシステムが、人種、性別その他の区別を含む個人に係るビッグデータに基づき統計的に何らかの評価・判断を行う場合（特に過去の実績データに基づいている場合）、当然ながら、その評価結果において、当該人種、性別等によって「差」が生じることになる。このような評価結果により一部のグループに不利益が生じる場合（例えば、犯罪捜査、人事採用など）、その判断結果が、人類が近代以降培ってきた、人種、性別等に関して人々は平等であるべきであるという「公平性、非差別性、平等性」という価値観（社会的規範）と反するではないかという問題が生じる。このような問題は、「私生活上の事柄をみだりに公開されない法的権利」というプライバシー権とは異なる新たな課題である。

さらには、これらの個人のIDに紐づく情報を国が管理することに関しても、市民スコアリングを含めて大きな議論がありうる。具体的には、国による社会全体の治安維持に加え、個別最適化され効率化された行政サービスの提供に資するのではないかとの考えもある一方で、不利益を被る立場からは国のサービスの公平性・平等性に反するとの指摘があることに加え、人間の「自由の確保」の観点からの、監視社会に対する批判があり、単なるプライバシーを超えた、基本的人権、民主主義の在り方など社会規範間の調整を必要とする課題であると言える。

これらの問題は、いずれも基本的は、AIそのものというよりも、むしろ個人データを扱うことによって生じる問題であるという側面も大きく、今後、少なくとも、プライバシーの保護という概念を超えた広義でのデータガバナンスの在り方が議論となると考えられる。

＜ガバナンスとしての透明性確保と説明可能性＞

この社会規範としての「人権・倫理」とは、「安全性」（生命・身体・財産の保護）と比較すれば、第３章（１）において記述したとおり、人類の歴史から見れば比較的新しい権利であり、今後とも時代によって中長期的に変化していくものであるとともに、国家体制など国・共同体によって、その重点の強弱は異なる[21]。また、それがゆえに、他の社会規範との優先に係る調整の考え方も含めて、その社会的な議論は必ずしも十分とは言えず、したがって詳細のルールは必ずしも確定・明確化されている訳ではない。

現在のAIシステムは、人間が与える一定のルールやデータに基づいて計算・判断するデジタル機械であり、したがって、そもそも明確なルールが確定されてないような人権・公平性に係る「倫理」を全て教え込み、判断させることには限界がある。

この観点から、とりわけ個人データを取り扱うようなAIシステムにおいては、そのガバナンスの一環として、より一層の「透明性」が求められる。すなわち、AIシステムに「倫理」に係る最終判断を任せることができないため、現時点では人・組織・社会に最終的に判断してもらうことが必要となる。具体的には、まずはシステムを運用する人・組織としての判断が必要となること（その際、組織の従業員においても多様性が求められること）に加え、さらには、個人・組織のみにおいても社会的に「正しい」とされる倫理を十分に判断することができないため、利害関係者その他を含む社会・共同体に判断・評価してもらい、社会的に受容されるべく、ガバナンス体制として透明性の確保が重要とされているものと考えられる。ただし、透明性が確保されれば、人権・公平性に係る「倫理」問題が全て解決するわけではないことに留意することが必要である。

この透明性の対象には、AIシステム全般が利用するデータの収取方法とその処理方法などの仕組みに加えて、場合によっては、当該AIシステムによる、個人の権利に影響を与えるような個別意思決定の理由に係る個人への開示なども含まれる。しかしながら、その際、トレーサビリティを確保する必要があることに加え、単にデータ・アルゴリズムが「透明化」されても、人々はAIシステムの結果を納得することができないという問題がある。これは、AI技術の中核となる機械学習技術は、ある意味でビッグデータの処理技術の一つにしか過ぎないものの、人間が自らプログラミングしたものではなく、いわゆるブラックボックスと言われる通り、必ずしも入出力の因果関係を含めたデータの処理の中身を人間の言葉に翻訳して説明することが容易にはできないという問題があるためである。このため、AIシステムの導入にあたっては、説明可能性という技術が必要となる。これは、広い意味で言えば、AIと人間が協調して作業を行うためのヒューマンインターフェースに係る技術であると言える。

（３）AI原則を支えるシステム・技術的課題

AIシステムが人類の有する社会規範・価値観と整合性を保ちつつ、社会の受容性を得て発展していくためには、図表３で示した技術・イノベーションの促進とガバナンスの構築との共進化の枠組みを考慮しつつ、イノベーションを促進するようなガバナンス体制を構築する

[21] 例えば、些細な話では、「年齢」による人事上の差別・区別は、米国では認められないが、日本では認められるなど。

とともに、社会規範に配慮したAIシステムに係る技術・システムの開発に取り組むことが必要である。

その際、（２）で示した通り、最終的にガバナンスに責任を持つ主体は人間であり、一方、AIシステムが複雑化する中で、ビッグデータにより作成されたAIシステムの挙動を、人間が必ずしも十分に把握できないという点が、現時点のAIにおける最も大きな技術的課題である。それに対応するための具体的な技術の事例として上述の（２）で述べた通り、AI信頼性（品質）評価手法、AIの説明可能性に係る技術があげられる。

- 信頼性（品質）評価：多量のデータを用い機械学習により生成されたAIは、人がルールベースで作ったプログラミングとは異なり、どの程度の品質が確保され所定の挙動範囲（フレーム）を超えるリスクがあるのかについて、人は必ずしも十分に把握できない。このため、品質評価手法に係る技術の確立が課題になる[22]。（日本では、産総研が先進的に取り組んでいる）
- 説明可能性：多量なデータを用いて機械学習により生成されたAIに関し、個々のAIでの判定結果について、人はその判定結果の理由について必ずしも理解できないという問題（いわゆるAIがブラックボックスと言われる側面）。これは、非常に多量のビッグデータに係る処理構造と、人間が言語で理解するロジックの処理構造にギャップが存在することによるものであり、このギャップを埋めるためにAIの説明可能性（Explainability）が重要になる。

すなわち、これらは、要は、「如何に人間がAIシステムの挙動を理解し、指示を出せるか」というヒューマンマシンインターフェースとの在り方であり、これを実現するためには、多量のデータ処理から情報を引き出す機械学習技術から、人間の思考ロジックを理解することができるAI技術への大きな技術的飛躍が必要である。このため、今後のAIの研究開発としては、「人との協調するAI」、さらには、人と対話して互いに向上していくような「人と共進化するAI」に向けた取り組みが重要になる。

また、上述に加えて、狭義のAI技術の進展に対応し、データ処理に係る基盤技術であるセキュリティ（教師データの改ざんへの対応など）、プライバシー（データ保護技術など）に加え、公平性（データの質・評価手法、公平性予測）を含めたデータガバナンスの体制を支える技術の開発も重要である[23]。

---

[22] 産業技術総合研究所は、世界に先駆けて機械学習の品質に係るガイドラインを発表している。
産業技術総合研究所「機械学習品質マネジメントガイドラインを公開－AIを用いた製品やサービスの品質を安全、安心に管理するために－」2020年6月30日
https://www.aist.go.jp/aist_j/press_release/pr2020/pr20200630_2/pr20200630_2.html

[23] なお、米国の国家AI研究開発戦略計画（2019リバイス）では、具体的な研究開発戦略（8戦略）として、①人間－AI協調（説明可能性など）、②ELSI（公平性など）、③安全性・セキュリティを挙げている（その他は、長期投資、データ、標準、人材、官民連携の横断的課題）。
US Whitehouse(Select Committee on Aritficial Intelligence of the National Science & Technology Councel) “THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE”2019年6月
https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf

５．AIガバナンスの限界と今後の社会影響

第４章まで、世界で策定されてきたAI原則をもとに、図表３に示す枠組みに沿って体系化することにより、社会規範を基にした、今後のデジタル・AIに対するガバナンスの方向を考察してきた。その際、基本的には、政府（国家）による、当該国に立地する組織に対するガバナンス構造を想定している[24]。

しかしながら、このようなガバナンス構造では、他の国家（第三国）に対して、自国におけるAIシステム利用に関し規律を要請することには限界が生じる。特に、当該第三国が、悪意・意図を持って当該国に対してAIシステムを利用しようとする場合には、国際的な合意による国際的なガバナンス体制を構築することも困難であろう。

また、自国内においても、政府（国家）自らによる当該AIシステムの利用に関しては、民主主義国家であれば、市民による監視メカニズムを構築することが可能であるが、そうでない国においては、適切なガバナンスの構築は困難であると考えられる。

具体的には、「AI倫理」の一部として常に議論される自律型致死性兵器システム（いわゆるAI兵器）の議論と、SNS等を利用した民主主義への影響行使があげられる。いずれにも、旧来型の戦争、国家間のサイバー攻撃・戦争に加え、「新たな戦争（国家間紛争）形態」として、今後の国際レジームや国家の在り方にも根本的な影響を与えかねない大きな問題であると言える。

- 「自律型致死性兵器システム（LAWS）」：LAWSの在り方については、現在国連の特定通常兵器使用禁止制限条約（CCW）の枠組みにおいて議論がなされている[25]。しかしながら、仮に規制しようにしても、技術の特定性・検証可能性の問題に加え、基本的には開発・利用主体が、国家（あるいは、ガバナンスの利かないテロ組織）などとなるため、規制・ガバナンスを強制する国際レジームの構築は容易ではないと考えられる[26]。

---

[24] なお、そもそも、AI技術は、汎用技術であり、多くの人が自由に扱うことができるとととともに、ソフトウェアであるため、技術的な封じ込めが困難であるという性質を有する。そのため、罰則などを設けても、その利用に係る国家による完全な監視や、特に悪用を完全な防止は困難であることに留意することが必要である。

[25] 福井康人（広島市立大学広島平和研究所元准教授）「自律型致死性兵器システム（LAWS）規制の動向」国際法学会エキスパート・コメントNo.2020-10、2020年6月9日
https://jsil.jp/archives/expert/2020-10

[26] 倫理的な観点からは、一般論としての「戦争反対」は、ほぼ全ての人が合意するものであり、人類の社会規範の一つであると言えるし、兵器の利用規制による戦争の回避については、多くの人が理想論として望ましいとの認識を有している。

しかしながら、現実の国際レジームの中で、仮に国家間の戦争は止むを得ないとした場合、自律的兵器の利用については、戦争に関係のない市民は極力巻き込まないなどとする戦争法（ジュネーブ法）の観点からは、より「倫理的」であり、望ましいのではないかという意見もありうる。一方で、人間が「ロボット」に殺されることは、倫理的に望ましくないという意見もありうる。

このLAWSについては、国連においても議論がなされているが、問題は、①むしろ（自律）兵器を作る能力のない国のみが禁止を主張する一方、（自律）兵器を作る能力のある国においては、少なくとも現時点においては、自律兵器の開発を止めるインセンティブが全く存在しないこと、また、②仮に国際的な開発・利用に係る規制に合意するとしても、規制を検証・行使するガバナンス体制の構築が困難であることにある。

- SNS等を利用した民主主義への影響行使：インターネット・SNSの個別利用者に対する（AIを活用した）マイクロターゲティングの活用については、今や商業活動（特に広告ビジネス）として多く利用されているが、その政治利用の在り方も含めて、民主主義体制への影響については、今後大きな課題である。特に、第三国・外国による特定の民主主義国に対する、選挙・世論などへの政治的な関与・攻撃については、現時点では、国内における言論・表現の自由の確保との兼ね合いもあり規制を行うことも難しく、また、国際的なガバナンスの在り方については、現時点においてほとんど議論がなされていない[27]。

特に、後者（②）については、そもそも致死型自律兵器の定義が困難であるため、特定することが困難である（すなわち例えば「自律」とは何かが規定できない）ことに加え、対象がソフトウェアであるため物理的に確認・検証が困難であることから（これらは、核技術とは異なる）、他国なり中立機関が信頼性をもって監査するようなガバナンススキームが非常に困難であるという問題である。このため、国際的な規制合意は、当面、非常に難しい状況にあると考えられる。

詳細は、例えば、以下を参照。

ポール・シャーレ「無人の兵団　AI、ロボット、自律型兵器と未来の戦争」早川書房（2019/7/18）（Paul Scharre“Army of None: Autonomous Weapons and the Future of War”, W. W. Norton & Company（2018/4/24））

[27] 「民主主義」は、世界各国を見渡すと、その成熟度等は国によって大きく異なるが、一般的には、人類共通の社会規範であると国際的にも概ね認識が共有されていると考えられる。この民主主義については、個々の個人（の大半）が、客観的情報に基づき合理的に判断するという前提に立っているが、インターネット・SNS、AIはこの前提に大きく影響を与えうる。

インターネット・SNSの利用は、2010年頃、アラブの春に見られる通り、従来、マスメディアが国家権力の影響を強く受けていた国において、当該マスメディアとは異なる客観的情報をリアルタイムで与えることを通じ、民主主義を促進するものとして、世界の注目を浴びた。

しかしながら、インターネット・SNSあるいはAIはツールにしかすぎず、誰がどのように使うかによって影響は大きく異なる。まず、SNSなどを通じて、AIの活用したマイクロターゲティングにより個人最適化される情報は、いわゆるエコーチェンバー現象を通じて、偏った内容のみが供給され、その結果、世論の二極化を引き起こすのではないかという指摘がある。

それに加えて、そのようなマイクロターゲティングのシステムを政治広告として利用することについては、民主主義の観点からどう評価するのかとの議論がある。特に、SNS等の利用者の心理・性格分析を含む個人情報に基づきターゲティングとして情報を提供すること、また、提供する情報が、明らかな政治広告としてだけではなく、例えば、一方の候補者のみに有利になるような感情のみ訴えるような記事や不完全あるいはフェイクなニュースを提供するという実態について、個人情報の保護、表現・言論の自由、他のメディアにおける広告規制との関係も含めて、その是非が課題になりうる。

特に、国家権力によるこのようなシステムの利用は、自らの都合の良い方向に活用することにより、当該国の民主主義に大きな影響を与える。特に、民主主義による政府に対するガバナンス体制が不十分な国においては、国家権力によるこのようなシステムの利用を抑えることができず、脱民主主義化が加速する可能性がある。さらには、中国に見られるとおり、国家権力が、国内の治安維持を目的に、国内におけるインターネット・SNS上の「有害」情報を制御することにより、国民の言論を抑えることも可能である。

さらに、ケンブリッジアナリティカの事例におけるロシアの関与疑惑の件に見られるとおり、外国（第三国）の機関が、このようなシステムを隠れて活用することにより、国内の世論や選挙結果が当該第三国にとって都合のよいように変えられるなど当該国の民主主義に大きな影響を与える可能性がある。これに対し、当該国においては、当該第三国に対してガバナンス上の対抗措置は有しない一

これまでの世界のAI原則において、前者のLAWSの問題については、非営利団体のFLIのアシロマ原則ではLAWSについて明示的に言及しているものの、ECのAI原則では懸念事例としては明記しつつも原則（要件）本体には明記されておらず、また、OECDのAI原則では全く明示されていない。後者の民主主義への影響に関しては、OECDのAI原則では、民主主義の価値の重要性に触れ、ECのAI原則では、民主主義の影響に係るインパクト評価を求めてはいるものの、いずれも一般論にとどまっており、具体的な方向については明示されていない。

このようなケースは、本ワーキングペーパーの図表３の枠組みに照らして考えると、技術・イノベーションの進展に対して、それらに抵触する社会規範に基づく要請に対して十分に機能するような制度・ガバナンスを構築することができないようなケースであると位置づけられる。

このように十分に機能するようなガバナンスを構築できない場合において、社会と技術・イノベーションはどのように発展していくのであろうか。また、何らかの調整を図るメカニズムが存在するのであろうか。近代以降の歴史を振り返ると、広義の情報通信技術である、印刷技術、電信・電話、無線技術・ラジオ・テレビなどの破壊的イノベーションは、封建体制から民主主義への移行を促し、また、一方で、人類の戦争の在り方そのものを根本的に変革してきた。近年のインターネット・スマートフォンなどに加え、新たなAIシステムに係る技術についても、今後そのガバナンス体制の構築が進む一方で、破壊的イノベーションとして、今後の世界の民主主義体制、国際紛争処理体制を含む国際レジームまで大きな影響を与える可能性があると考えられるが、具体的にどのような影響を与えるのかは、更なる研究・考察が必要であり、今後の検討課題である。

（以上）

方、国内では表現・言論の自由の確保のため、当該システムの利用に規制を導入することが困難になるなど、難しい状況に陥ることが想定される。

関連する書籍としては、例えば、以下を参照。

P・W・シンガー、エマーソン・T・ブルッキング他、「「いいね！」戦争 兵器化するソーシャルメディア」NHK出版（2019/6/20）

NHK取材班「AI vs.民主主義: 高度化する世論操作の深層」NHK出版（2020/2/10）

ジェイミー・バートレット、「操られる民主主義　デジタル・テクノロジーはいかにして社会を破壊するか」草思社（2018/9/25）

笹原 和俊、「フェイクニュースを科学する: 拡散するデマ、陰謀論、プロパガンダのしくみ」化学同人（2019/2/27）

（参考１）各国政府、国際機関等におけるAI原則を巡る経緯

（１）日本のAI原則

＜総務省の動きと国際的な働きかけ＞

筆者の理解する限りにおいて、世界各国の政府においてAI原則を作ろうという議論の始まりは、日本の総務省情報通信政策研究所である。

（報告書2015）

総務省情報通信政策研究所は、2015年1月に、「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」の開催を発表した[28]。同発表においては、

> *「2045年にはコンピュータの能力が人間を超え、技術開発と進化の主役が人間からコンピュータに移る特異点(シンギュラリティ)に達するとも議論されるなど、その処理能力は加速度的に高まっています。」*
>
> *「ビッグデータ、人工知能、ロボット等を通じて、既に私たちはこれら技術の恩恵を受け始めています。しかしこれらは始まりであって、十年後、二十年後には、今の私たちにはSFとも思われる世界が広がっている可能性があります。」*

とし、ICTインテリジェント化がもたらす可能性、社会へのインパクト、政策課題などについて検討することとした。

５回の審議を経て、2015年6月にまとめられた報告書（「報告書2015」）[29]では、「インテリジェントICTの研究・開発に係る原則の検討」が提言されている。同「原則の検討」においては、Issac Asimovのロボット３原則[30]を参考にするとともに、制御可能性、セキュリティ、プライバシー保護などが項目として考えられるとし、また、The Future of Life Instituteの公開質問状などの動きについても触れている。

（報告書2016）

---

[28] 総務省情報通信政策研究所「「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」の開催」、2015年1月27日
https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000024.html

[29] 総務省情報通信政策研究所「「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2015」2015年6月30日
https://www.soumu.go.jp/main_sosiki/kenkyu/intelligent/index.html

[30] 第一法則：ロボットは人間に危害を加えてはならない．またその危険を看過することによって，人間に危害を及ぼしてはならない。
第二法則：ロボットは人間から与えられた命令に服従しなくてはならない．ただし，与えられた命令が第一法則に反する場合はこの限りではない。
第三法則：ロボットは前掲の第一法則，第二法則に反するおそれのない限り，自己を守らなければならない

その半年後の2016年1月、総務省情報通信政策研究所は、「ICTインテリジェント化影響評価検討会議」の開催を発表し[31]、上記報告書2015を踏まえて、ICTインテリジェント化に関し、目指すべき社会像及びその基本的理念を検討するとともに、インパクトスタディ・リスクスタディを行うこととした（なお、同会議は、3月の第二回会合において、「AIネットワーク化検討会議」と名称変更がなされている）。

同会議は、2016年4月、中間報告書「AIネットワーク化が拓く智連社会（WINS）ー第四次産業革命を超えた社会に向けてー」を発表した[32]。同中間報告書においては、OECDプライバシーガイドライン、同セキュリティガイドライン等を参考に、国際的に参照される枠組みとして研究開発の原則を策定すべく検討に着手するものとし、少なくとも、８つの原則、すなわち、①透明性の原則、②利用者支援の原則、③制御可能性の原則、④セキュリティ確保の原則、⑤安全保護の原則、プライバシー保護の原則、⑦倫理の原則、⑧アカウンタビリティの原則、を盛り込むべきとしている。

中間報告書発表後の2016年4月、日本の香川県において、G7香川・高松情報通信大臣会合が開催された[33]。同共同宣言には、特段AI原則については明示的に触れられていないものの、上記8原則案を配布したとのことであり、総務省の発表資料[34]によると、「我が国より、AIネットワーク化が社会経済に与える影響の分析を国際機関も含めた連携を通じて実施し、AIの開発原則の議論へとつなげていくことを提案した」としており、「高市総務大臣からの提案に対し、各国から賛同が得られた」としている。

2016年6月にまとめられた同会議の最終報告書「報告書2016」[35]においては、上記G7会合の報告を記載するとともに、上記８原則に関し、ブレーク7ダウンをするとともに、国際社会に向けて、OECD等における継続的議論の必要性を提唱している。

（報告書2017、報告書2018、報告書2019）

総務省情報通信政策研究所は、2016年10月から新たにAIネットワーク社会推進会議を開催し、2017年6月に「報告書2017」の案をまとめたあと、パブコメを経た上で、同年7月に報告書を確定、公表している[36]。「報告書2017」においては、「国際的な議論のためのAI開発ガイドライン案」が掲載され、上記８原則に、連携の原則を加えた、９原則について、その解説も含めてガイドラインとして提示している。

総務省は、上記ガイドライン案を検討中の2016年11月に、OECDで初めて開催されたAI関連の会議である「AIに関するOECD技術予測フォーラム2016」に、上記検討状況をインプットす

---

31 総務省情報通信政策研究所「ICTインテリジェント化影響評価検討会議の開催」2016年1月27日
https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000039.html

32 総務省情報通信政策研究所「「AIネットワーク化検討会議」中間報告書の公表」2016年4月15日
https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000049.html

33 総務省「G7香川・高松情報通信大臣会合」2016年4月29日～30日
https://www.soumu.go.jp/joho_kokusai/g7ict/index.html

34 総務省「G7香川・高松情報通信大臣会合の開催結果」2016年4月30日
https://www.soumu.go.jp/menu_news/s-news/01tsushin06_02000083.html

35 総務省情報通信政策研究所AIネットワーク検討会議「報告書2016：AIネットワーク化の影響とリスク - 智連社会（WINS）の実現に向けた課題」2016年6月20日
https://www.soumu.go.jp/main_sosiki/kenkyu/iict/index.html

36 総務省情報通信政策研究所AIネットワーク推進会議「報告書2017：AIネットワーク化に関する国際的な議論の推進に向けて」2017年7月28日
https://www.soumu.go.jp/iicp/research/results/ai-network.html

るとともに、上記ガイドライン案が確定したあとの2017年10月に、総務省の支援の下で、パリにてOECDのAIに係る会議「AI：知能機械、スマートな政策」を開催し[37]、同会議において、上記９原則について報告を行っている[38]。

なお、2017年9月にイタリアにて開催されたG7情報通信・産業大臣会合[39]においては、AIに係る付属文書（「我々の社会のための人間中心のAI社会に関するマルチステークホルダーの交流」）が添付され、アカウンタビリティ、透明性、プライバシー、サイバーセキュリティ、安全性に係る議論の必要性に触れている。しかしながら、原則について触れているわけではない。

なお、総務省その後引き続きAIネットワーク推進会議を開催しており、2018年7月には「報告書2018」を公表し、AI利活用原則案（10原則）を提示するとともに、2019年8月には「報告書2019」を公表し、AI利活用ガイドライン（10原則）を発表している[40]。

＜内閣府の動き＞

一方、上記の総務省の動きに加え、日本政府全体として、2019年に予定されているG7（@フランス）、G20（@大阪）に向けて原則論を打ち出していくべきとの議論の中で、内閣府において、2018年から「AI社会原則」の検討が始まっている。

まずは、内閣府では、日本・香川でのＧ7情報通信大臣会合のあとの、2016年5月に、「人工知能と人間社会に関する懇談会」を立ちあげ、2017年3月に報告書をとりまとめている[41]。本報告書では、倫理、法令、経済、教育、社会、研究開発の観点からそれぞれ論点をまとめているが、特段「原則」に触れている訳ではない。

その上で、再度、2018年4月に、内閣府の主導の下、人工知能技術戦略会議の下に25名の専門家からなる「人間中心のAI社会原則検討会議」[42]を設置し、翌月から検討を開始している。同会議では、総務省のAI開発ガイドライン案よりも上位の「社会原則」を議論することとし、G7及びOECD等の国際的な議論に供することをその目的の一つとして掲げている。

同会議は、2018年12月に「人間中心のAI社会原則（案）」を公表し、その後パブコメを経た上で、2019年3月29日に、同原則を統合イノベーション戦略推進会議決定し、公表してい

---

37 OECD Conference on Artificial Intelligence - "AI: Intelligent Machines, Smart Policies", 2017年10月26 - 27日
https://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/

38 なお、会議のサマリーでは、原則そのものには記載はないものの、適切なセーフガードの必要性に加えて、透明性と監視、アルゴリズムによる差別、プライバシー、責任、セキュリティ・安全性の問題を掲げている。

39 総務省「G7情報通信・産業大臣会合の開催結果」2017年9月26日
https://www.soumu.go.jp/menu_news/s-news/01tsushin06_02000103.html
http://www.g7italy.it/en/documenti-ministeriali/index.html

40 総務省AIネットワーク社会推進会議 「報告書2018の公表－AIの利活用の促進及びAIネットワーク化の健全な進展に向けて－」、2018年7月17日
総務省AIネットワーク社会推進会議「報告書2019」、2019年8月9日
https://www.soumu.go.jp/iicp/research/results/ai-network.html

41 内閣府「人工知能と人間社会に関する懇談会」
https://www8.cao.go.jp/cstp/tyousakai/ai/index.html

42 内閣府「人間中心のＡＩ社会原則検討会議」
https://www8.cao.go.jp/cstp/tyousakai/humanai/index.html
なお、同会議は2019年2月に、内閣官房統合イノベーション戦略推進会議の下に移行されている。
https://www.cas.go.jp/jp/seisaku/jinkouchinou/

る。同原則では、AI社会原則として、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、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の７つの原則を提示している。

＜日本人工知能学会の動き＞

一方、日本の人工知能学会（JSAI）は、比較的早い段階から、いわゆるAIの倫理問題に取り組んでいる[43]。

同学会は、2014年9月に理事会の承認を受けた上で、同年12月に倫理委員会創設した。当時の委員長である松尾氏は、別の名称も検討したが、「倫理委員会」という名称は、（AIが）何らかの影響力をもつ「怖い」委員会であるというイメージがあり、社会と対話して進めていくという観点から決定したと振り返っている[44]。

その後、2015年6月の公開討論会、2016年6月の公開討論会：案を公開・議論を経て、2017年2月に理事会の承認を受けて、「倫理指針」を発表した[45]。同指針では、①人類への貢献、②法規制の遵守、③他社のプライバシーの尊重、④公平性、⑤安全性、⑥誠実な振る舞い、⑦社会に対する責任、⑧社会との対話と自己研鑽、⑨人工知能への倫理遵守の要請の９項目からなる。

（２）欧州の動き

欧州では、それまでに学会、民間レベルの動きはいくつかあったものの、欧州全般として、政府のAI政策を打ち出すのが比較的に遅かったことを踏まえ、欧州委（EC）におけるAI原則の動きも、2018年以降と比較的に遅くスタートしている。しかしながら、その後積極的な取組を進め、新たな規制の導入も検討している段階にある。

欧州委（EC）の独立諮問機関である「科学と新技術での倫理に関する欧州グループ（EGE）」は、2018年3月に、「人工知能、ロボット及び自律システムに係る声明」を発表した[46]。同声明においては、EU条約及びEU基本的人権憲章に記載された価値観を踏まえて、基本的な倫理原則を策定することを提案するとともに、倫理原則として、①人間の尊厳、②自律性、③責任、④正義・平等・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・物理的精神的統合性、⑧データ保護とプライバシー、⑨持続可能性を例示している。

なお、同声明においては、これまでの他の組織の動きとして、IEEE、ITUや学会（ACM/AAAI）、民間企業（IBM、マイクロソフト、グーグル、PAI）などの動きに加え、特にFuture of Lifeのアシロマ原則やモントリオール大学の宣言（2017年10月の第一版）について言及している。なお、日本の動きについては特段触れられていない。また、同声明における問題意識としては、特に「自律性」における人間の尊厳や道徳的責任との関係に言及しており、具体

---

[43] 人工知能学会倫理委員会
http://ai-elsi.org/about

[44] 人工知能学会倫理委員会設立の趣旨
http://ai-elsi.org/about/purpose

[45] 「人工知能学会　倫理指針」について
http://ai-elsi.org/archives/471

[46] European Group on Ethics in Science and New Technologies “EGE Statement on artificial intelligence, robotics and 'autonomous' systems 2018”、2018年3月9日
https://ec.europa.eu/info/publications/ege-statements_en

的には、自動運転（トロッコ問題を超えた議論）、自律兵器、自律ソフトウェア（Siri、Alexaなど。ボットを含む）を問題視している。

これを踏まえて、欧州委（EC）は、2018年4月に発表したAI戦略である「欧州のためのAI」[47]では、EGEと協力しつつ、年内までに基本的権利に係るAI倫理ガイドラインのドラフトを開発する旨記載された。それを踏まえて、ECは、同年6月に、52名の専門家からなるAIに係るハイレベル専門家グループ（High-Level Expert Group on AI：AI-HLEG）を設置した。同グループは2018年12月に「信頼できるAIに向けた倫理ガイドライン・ドラフト案」[48]を発表し、その後関係者との議論、500以上のコメントの受付を経た上で、2019年4月、同ガイドラインの確定版を発表している[49]。

同ガイドラインでは、信頼できるAIの要件として、①人間の代理・監督、②技術的頑健性・安全性、③プライバシー・データガバナンス、④透明性、⑤多様性・被差別・公平性、⑥社会的・環境的幸福、⑦説明責任の７要件を掲げている[50]。また、AIによる得られる機会の例示として、気候変動対応、健康・幸福、質の高い教育とデジタルフォーメーションを挙げる一方、深刻な懸念の例として、個人の特定とトラッキング、隠れた（Covert）AIシステム（人間がAIと交流する際、相手がAIであるということ気づかせることが必要）、基本的人権に違反する市民スコアリング、自律兵器システム（LAWS）の４つの事例を挙げている。

なお、欧州委（EC）は、2019年4月の上述のガイドラインの発表と併せて、その次の取組方針として、「人間中心AIにおける信頼性構築」を発表している[51]。同方針においては、本ガイドラインに記載された評価リストをもとに、関係者に実際に試してもらいフィードバックを行い、2019年末までに再度評価を行い、2020年初にはガイドラインをアップデートするとしている[52]。また、同方針では、国際的なAI倫理ガイドラインの作成に向けて、EUのアプローチを世界に広げるべく、G7，G20を含め国際的な議論に積極的に関与するとしている。

---

[47] European Commission COM(2018)237 Final “COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Artificial Intelligence for Europe{SWD(2018) 137 final}, 2018年4月25日
https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[48] The European Commission’s HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE “DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI” Working Document for stakeholders’ consultation Brussels, 18 December 2018
https://ec.europa.eu/digital-single-market/en/artificial-intelligence

[49] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE “ETHICS GUIDELINES FOR TRUSTWORTHY AI”, 2019年4月
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

[50] なお、2018年12月のドラフト版では、1. Accountability, 2. Data Governance, 3. Design for all, 4. Governance of AI Autonomy (Human oversight), 5. Non-Discrimination, 6. Respect for (& Enhancement of) Human Autonomy, 7. Respect for Privacy, 8. Robustness, 9. Safety, 10. Transparencyの10要件となっており、セット版では大きく変更されている。

[51] European Commission COM(2019) 168 final COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Building Trust in Human-Centric Artificial Intelligence, 2019年4月8日
https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

[52] 2020年9月時点で、まだアップデートは発表されていない模様である。

また、欧州委（EC）は、2020年２月に「AI白書」を発表した[53]。同白書においては、AIの「信頼」に係るエコシステムを作るべく、既存の法的枠組みの調整に加え、AIに係る特別な法制度やAIに係る欧州の統治機構の必要性を提議している。

（３）米州、民間団体の動き

＜米国連邦政府の動き＞

米国では、いわゆるAIの倫理・社会問題に関し、オバマ政権時に先進的に取り組んでいたが、トランプ政権時になって動きが全く止まった。その後、トランプ政権は、2019年にAI戦略を発表したが、AIに対する規制に関しては、非規制的アプローチを強く志向している。

オバマ政権は、2016年5月から検討を開始、大学・NPO等とともに４回のワークショップを開催した後、2016年10月に「人工知能の未来に備えて」を発表した[54]。同報告書では、AIと規制、研究と労働力に加え、公正・安全・ガバナンスなどについて記述がなされ、23の提言の記載がなされているが、AI原則などの必要性ついては、特段触れられていない。

その後、2017年からのトランプ政権においては、特段動きがなかったが、2019年2月になって、「人工知能での米国のリーダーシップの維持に関する大統領令（AIイニシアティブ」を発表[55]し、同イニシアティブにおいては、OMBは180日以内に各省庁向けのAI応用に係る規制のガイダンスに係るメモランダムを発行することが求められた。これを踏まえ、OMBは、2020年1月に同メモランダムを発行[56]し、意見聴取を開始した。同メモランダムでは、非規制的アプローチを強く志向（どうしても必要と判断したときのみ、新たな規制を検討）するとともに、各省庁が規制・非規制的アプローチを検討する際に遵守すべき10の原則を提示ししている[57]。

なお、米国国防総省（DOD）は、2018年6月に省内に共同AIセンター（JAIC）を設置し、2019年2月に同省としてのAI戦略を発表しているが、2020年2月には同省としてのAI倫理原則を

---

[53] EUROPEAN COMMISSION, COM(2020) 65 final, “WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust”, 2020年2月19日
https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[54] Executive Office of the President, National Science and Technology Council Committee on Technology, “PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE”, 2016年10月
https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[55] Executive Order on Maintaining American Leadership in Artificial Intelligence, 2019年2月11日
https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[56] 米国OMB, “MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES, Guidance for Regulation of Artificial Intelligence Applications”, 2020年1月7日
https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

[57] Principles for the Stewardship of AI Applicationsとのタイトルの下で、①AIに係る国民の信頼、②国民の参加、③科学的統合性と情報の質、④リスクアセスメントとマネジメント、⑤便益とコスト、⑥柔軟性、⑦公平性と非差別、⑧公開と透明性、⑨安全性とセキュリティ、⑩省庁間調整の項目からなる。

採択している[58]。同原則は、責任あること、公平であること、追跡可能であること、信頼できること、ガバナンス可能なことの５項目からなり、JAICが中心となって運用することとしている。

＜The Future of Life Institute（FLI）＞

Future of Life Institute（FLI）は、MIT教授その他によって2014年5月に設立された非営利機関であり、Elon Mask氏が多額の寄付をするとともに、物理学者のスティーブン・ホーキング博士、未来学者のレイ・カーツワイル博士などが支持者として参加していることで有名。同団体は、2015年以降、AIの未来に係る各種会議を開催し[59]、研究優先事項に係る公開質問状[60]を発表するなどの活動を行っている。

FLIは、2017年1月、米国カリフォルニア州のアシロマにて会議を開催し、翌月アシロマAI原則を発表した[61]。同原則は、23の原則（研究課題5原則、倫理・価値13原則、長期的課題5原則）からなる。うち、倫理・価値に係る項目は、安全性／障害の透明性／司法の透明性／責任／価値観の調和／人間の価値観／個人のプライバシー／自由とプライバシー／利益の共有／繁栄の共有／人間による制御／非破壊／人工知能軍拡競争からなる。なお、この「アシロマ」は、1975年に遺伝子組み換え技術にガイドラインが議論され、生物学的封じ込めの合意がなされた「アシロマ会議（Asilomar conference）」の場所として有名。

なお、2020年9月現在、1677名のAIロボット研究者、3662名のその他の方々が本原則に署名しているとのこと。

＜Partnership on AI（PAI）＞

Partnership on AI（PAI）は、2016年9月に、AIの責任ある利用にコミットすべく創設された非営利団体であり、設立時のメンバーは、アマゾン、フェースブック、グーグル、Deep Mind、マイクロソフト、IBMの６社。現在は、100以上の団体が参加しており、その約６割が非営利機関、約２割が大学等、約２割が企業。

PAIは、設立当初において、８項目からなるTenets（教義）[62]を発表している。その内容は、便益・能力向上、市民の聴取等、ステークホルダーとの連携、倫理問題等への開かれた研究と対話、研究開発への取組などに加え、プライバシー、セキュリティ、研究者などの社会的責任等にも触れている[63]。

---

58 DOD Adopts Ethical Principles for Artificial Intelligence, 2020年2月24日
https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

59 例えば、第一回の会議は、2015年1月「The Future of AI: Opportunities and Challenges」。
https://futureoflife.org/2015/01/11/ai-conference/

60 「公開質問状　堅牢かつ有益な人工知能のための研究優先事項」
https://futureoflife.org/ai-open-letter-japanese/

61 Future of Life Institute「アシロマAI原則」（2017年2月3日）
https://futureoflife.org/ai-principles/
https://futureoflife.org/ai-principles-japanese/

62 Partnership on AI: Tenets
https://www.partnershiponai.org/tenets/

63 具体的には、①可能な限り多くの人に便益・能力向上、②市民の教育・聴取、ステークホルダーとの連携、③倫理的、社会的、経済的、法的関係について開かれた研究と対話、④幅広いステークホルダーに対する説明責任の下で、研究開発への積極的な取り組み、⑤分野別の懸念・機会の理解を確

＜IEEE＞

IEEE（Institute of Electrical and Electronics Engineers）は、早い段階から、AI及び自律システムに係る問題に取り組んでいる。

具体的には、「倫理的に揃えられたデザイン（Ethically Aligned Design: EAD）～自律・知能システム（A/IS）の人間の幸福への重点化に向けたビジョン」のレポートに取り組んできており、2016年12月にVersion1発表し、パブコメ等を経たうえで、2017年12月にVersion2発表した[64]。その後、さらにパブコメ等を経たうえで、2019年6月、First Editionを理事会承認、公表している[65]。

同レポートには、「一般原則」という内容があるが、バージョンごとに内容が追加されてきている。具体的には、Version 1では4原則（Human Benefit, Responsibility, Transparency, Education and Awareness）、Version 2では5原則（ Human Rights, Prioritizing Well-being, Accountability, Transparency, A/IS Technology Misuse and Awareness of It）となっていたが、First Editionでは、人権、幸福、データエージェンシー、効率性、透明性、説明責任、誤用に係る気づき、能力の８つの原則が記載されている。

＜カナダ（モントリオール大学）＞

モントリオール大学は、ケベック州研究基金との連携のもとで、2017年11月に、プロジェクトの開始の表明と「AIの責任ある開発に関する宣言」第一版を発表している。その後、G7のAIにマルチステークホルダー会議がモントリオールにて開催された2018年12月に、「AIの責任ある開発に関するモントリオール宣言」として最終的に発表している[66]。

具体的には、10の原則、8の勧告から構成され、うち、原則としては、①幸福、②自律性尊重、③プライバシーと親密性の保護、④連帯、⑤民主的参加、⑥平等性、⑦多様性包摂性、⑧慎重性、⑨責任、⑩持続的発展の10の原則が記載されている。

なお、2020年8月現在、1932名の市民、108の機関が署名しているとのことであり、カナダの連邦政府も参加している。

（４）アジア（中国、シンガポール）の動き

＜中国＞

中国では、下記G20が開催される直前に、AI原則に取り組んでいる。具体的には、AI担当の横断的部署である「次世代AI発展計画推進室」が、2019年3月に、「国家次世代AIガバナ

保すべく、ビジネス界のステークホルダーと連携、⑥プライバシー・セキュリティの保護、影響を受ける人の関心の理解・尊重、研究者・技術者の社会的責任等の確保、技術の頑健性・信頼性・安全条件下での運用、国際条約・人権に反する技術の利用に反対、⑦技術の説明の目的のための理解可能、解釈可能性の重要性を認識、⑧科学者・技術者間での協力・信頼・開放性の文化の創出

[64] IEEE Ethically Aligned Design, Version 1, Version 2
https://standards.ieee.org/industry-connections/ec/ead-v1.html

[65] IEEE, Ethically Aligned Design, First Edition
https://ethicsinaction.ieee.org/

[66] The Montreal Declaration for the Responsible Development of Artificial Intelligence Launched
https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/
https://www.montrealdeclaration-responsibleai.com/

ンス専門委員会」を設置し、その３か月後の2019年6月17日に「次世代AIガバナンス原則 －責任を有するAIの発展」を発表している[67]。

内容は8項目であり、①調和（和諧）・友好、 ②公平・公正、③包摂・共有、④プライバシーの尊重、⑤セキュリティ・制御可能性、⑥責任の分担、⑦開放・協力、⑧アジャイルガバナンスとなっている。

＜シンガポール＞

シンガポールは、AI原則の内容を深く検討するというよりは、ガバナンス体制の構築について他国に先んじて取り組んでいる[68]。

シンガポールの情報通信メディア開発庁（IMDA）は、2018年6月、「AIガバナンスと倫理イニシアティブ」という計画を発表[69]。同計画では、AIの倫理基準を設けるべく、IMDAの下に諮問委員会を設置するとともに、個人情報保護委員会（PDPC）がディスカッションペーパーを発表、また、AIガバナンスとデータ利用に係る研究プログラムを開始するとしている。なお、同ディスカッションペーパーでは、①説明可能性・透明性・公平性、②人間中心的、の２点を鍵となる原則としている。

シンガポール政府は、2019年1月、モデルAIガバナンス枠組み（第一版）を発表[70]。その後、さらに意見を聴取、検討を重ねた上で、2020年1月に同第二版を発表[71]。同第二版では、ガバナンスの枠組みとして①内部ガバナンス構造、②AI意思決定における人間関与のレベルの決定、③運営マネジメント、④利害関係者との交流・コミュニケーションをあげている。なお、参考資料として「既存のAI原則」との説明の下、９つの原則を例示している[72]。

また、併せて、同枠組みと整合した「組織のための実行・自己評価ガイド（ISAGO）」を発表している[73]。これは世界経済フォーラム第四次産業革命センター（WEF C4IR）と連携し

---

67 総務省「国内外の議論及び国際的な議論の動向」2019年7月
https://www.soumu.go.jp/main_content/000646182.pdf

68 日経XTECH「「AIと倫理」に一石、シンガポールの戦略」江間有沙、2018年8月23日
https://xtech.nikkei.com/atcl/nxt/column/18/00412/082100001/

69 ARTIFICIAL INTELLIGENCE GOVERNANCE AND ETHICS INITIATIVES
https://www.imda.gov.sg/-/media/imda/files/about/media-releases/2018/2018-06-05-fact-sheet-for-ai-govt.pdf?la=en

70 Model AI Governance Framework
https://www.imda.gov.sg/AI

71 MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK SECOND EDITION
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

72 アカウンタビリティ、正確性、監査可能性、説明可能性、公平性、人間中心と幸福、人権との整合性、包摂性、先進性。

73 Companion to the Model AI Governance Framework - Implementation and Self-Assessment Guide for Organizations
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf

て作成したもの[74]であり、60以上の機関として作成したとのこと。また、併せて、Compendium（大要）として、８つのケーススタディーを記載した資料も発表している[75]。

（５）国際機関（OECD、UNESCO等）の動き

＜OECD・G20の動き＞

OECDにおけるAIに係る取組としては、2016年11月に、OECDで初めて開催されたAI関連の会議である「AIに関するOECD技術予測フォーラム2016」を、また、2017年10月に、日本の総務省の支援の下で、パリにてOECDのAIに係る会議「AI：知能機械、スマートな政策」を開催しており、総務省は、同会議において、日本のAI開発原則案（９原則）について報告を行っている。

OECDのCDEP（デジタル経済政策委員会）は、2017年11月、2018年5月に会合を開催し、2019年以降の理事会勧告（AI原則を含む）作成に向けた作業に着手することで合意した[76]。具体的には、OECD/CDEPは、2018年5月、原則策定のための専門家グループの創設に合意。50名以上の専門家からなるAIGO（AI Group of experts at the OECD）を創設し、2018年9月～2019年2月にかけて、4回の会合を実施（UAEでの会合含む）。2019年3月、CDEPは、最終ドラフトを了承。2019年5月、閣僚会議にて「AIに係る理事会勧告」を採択[77]。

同勧告には、「信頼できるAIのための責任あるスチュワードシップに係る原則」が含まれ、①包摂的な成長、持続可能な開発及び幸福、②人間中心の価値観及び公平性、③透明性及び説明可能性、④頑健性、セキュリティ、安全性、⑤アカウンタビリティの５つの項目から構成されている。

なお、このOECDのAI原則とほぼ同様の内容がG20でも合意されている。具体的には、2019年6月、G20茨城つくば貿易・デジタル経済大臣会合が開催[78]されたが、その同閣僚声明において「G20 AI原則」を添付された。また、本資料は、同月開催されたG20大阪首脳会議においても、付属文書として添付されている[79]。

---

[74] Singapore and World Economic Forum driving AI Adoption and Innovation、2020年1月22日
https://www.imda.gov.sg/news-and-events/Media-Room/Media-Releases/2020/Singapore-and-World-Economic-Forum-driving-AI-Adoption-and-Innovation

[75] Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGAIGovUseCases.pdf

[76] 総務省「報告書2018」、2018年7月17日、より。
https://www.soumu.go.jp/main_content/000564157.pdf
総務省「国際的な議論及び海外の議論の動向」2018年2月21日、より
https://www.soumu.go.jp/main_content/000537165.pdf

[77] OECD “Recommendation of the Council on Artificial Intelligence”, 2019年5月22日
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[78] 総務省「G20茨城つくば貿易・デジタル経済大臣会合の開催結果」2019年6月11日
https://www.soumu.go.jp/menu_news/s-news/01tsushin08_02000106.html

[79] 外務省「G20 大阪サミット」2019年6月28日～29日
https://www.mofa.go.jp/mofaj/gaiko/g20/osaka19/jp/documents/

＜UNESCOの動き＞

国連（UNESCO）では、2018年頃からAIに係る取組を進めている[80]が、2019年3月に、日本の支援の下で、AI原則に係る国際会議をパリにて開催した[81]。その後、2019年4月の執行委員会での決定を経て、2019年11月に第40回全体総会において、AIの倫理に関する勧告を2021年の総会で採択することを目指すことを決定した。

これを踏まえて、UNESCOは、2020年5月にAI倫理に関する勧告一次案を発表[82]し、同年7月からオンラインでのパブリックコンサルテーションを開始している。同一次案では、６つの価値を示すとともに、12の原則（「人間と繁栄のために、プロポーショナリティ、人間の監督と決定、持続可能性、多様性と包摂性、プライバシー、意識とリテラシー、マルチステークホルダーと適応的ガバナンス」「公平性、透明性と説明可能性、安全性とセキュリティ、責任と説明責任」）が示されている。

（以上）

[80] UNESCO Artificial intelligence with human values for sustainable development
https://en.unesco.org/artificial-intelligence

[81] UNESCO Conference ″Principles for AI: Towards a Humanistic Approach?
https://en.unesco.org/events/unesco-conference-principles-ai-towards-humanistic-approach
日本経済新聞「AI倫理の報告書作成へ　ユネスコ、議論けん引図る」2019年3月5日
https://www.nikkei.com/article/DGXMZO42030560V00C19A3CR0000/

[82] UNESCO: first version of a draft text of a recommendation on the Ethics of Artificial Intelligence
https://unesdoc.unesco.org/ark:/48223/pf0000373434

（参考２）主なAI原則におけるAIの定義

機関	定義
OECD	AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.
EU	Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
IEEE(A/IS)	We prefer not to use-as far as possible-the vague term “AI” and use instead the term autonomous and intelligent systems (A/IS). This terminology is applied throughout Ethically Aligned Design, First Edition to ensure the broadest possible application of ethical considerations in the design of the addressed technologies and systems.
UNESCO	AI systems embody models and algorithms that produce a capacity to learn and to perform cognitive tasks, like making recommendations and decisions in real and virtual environments. AI systems are designed to operate with varying levels of autonomy by means of knowledge modeling and representation and by exploiting data and calculating correlations. AI systems may include several approaches and technologies, such as but not limited to: i. machine learning, including deep learning and reinforcement learning, ii. machine reasoning, including planning, scheduling, knowledge representation, search, and optimization, and iii. cyber-physical systems, including internet-of-things and robotics, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.
総務省	「ＡＩ」とは、「ＡＩソフト及びＡＩシステムを総称する概念」をいう 。 ●「ＡＩソフト」とは、データ・情報・知識の学習等 により、利活用の過程を通じて自らの出力やプログラムを変化させる機能を有するソフトウェアをいう。例えば、機械学習ソフトウェアはこれに含まれる。 ●「ＡＩシステム」とは、ＡＩソフトを構成要素として含むシステムをいう。例えば、ＡＩソフトを実装したロボットやクラウドシステムはこれに含まれる。