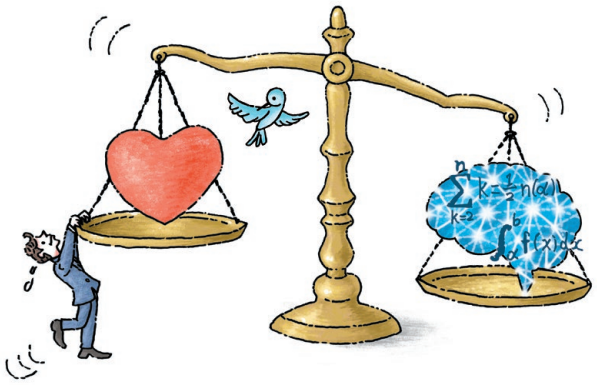
情報通信の現在と未来を展望する

Vol.50
2022 Summer
ネクストコム

Nextcom

特集

AI倫理とガバナンス



Feature Papers

特集論文
AIガイドライン、トラスト、ガバナンス
—最適なルール策定に向けて—
実積 寿也　中央大学 総合政策学部 教授

特集論文
人工知能とリスク分析文化
久木田 水生　名古屋大学 大学院 情報学研究科 准教授

特集論文
AIガバナンスを支える技術
黒川 茂莉　株式会社KDDI総合研究所 AI部門
麻生 英樹　株式会社KDDI総合研究所 AI部門
国立研究開発法人 産業技術総合研究所 人工知能研究センター

Articles

5年後の未来を探せ
気候変動を左右する海洋の微量金属元素を追う
近藤 能子　長崎大学 大学院 准教授に聞く
江口 絵理 ライター

明日の言葉

よりよくする方法がある。それを見つけよ。
……トーマス・エジソン

エジソンが自分の研究室に掲げていたという言葉。
その後、ニューヨークタイムズなどにも、引用句が掲載された。

INDEX
summer
2022

特集

AI倫理とガバナンス

写真：道を探しつつ歩むラクダの隊列。
シルクロードの地、トルファン（中国新疆ウイグル自治区）。

すでに始まってしまった未来について —— ⑤⓪

文：平野啓一郎
絵：大坪紀久子

「言葉が通じる外国」

十年ほど前、ニューカレドニアの国際文学シンポジウムに参加したことがある。どうしてニューカレドニアで？と最初は不思議だったが、フランス本国の文化政策で、フランス文化の普及に資する事業には助成金が出ているらしい。フランス語を話せることが参加の条件で、私のフランス語はお粗末だが、それでもどうにかパネルをこなし、地元のラジオにまで出演した。良い思い出である。

フランスからの移民は、二世三世が多いのだろうと思い込んでいたが、意外に一世の人たちと多く出会った。私はつい、どうして移住してきたのかと理由を尋ねたが、二人に尋ねたところで以後はそれを慎んだ。彼らはいずれも、少し言い淀んでから、「ヨーロッパが嫌になった」といった、ほとんどゴーギャンのような理由を口にして詳細を語らなかった。

現地の大学の教授から聞いた話では、本国から移住してくる人は色々で、ニューカレドニアの自然に惹かれてくる人もいれば、諸事情で本国に居辛くなった人もいるとのことだった。

私はその時、日本にもし、こんな風な日本語が通じる本土以外の土地があったならば、どうだっただろうかと想像した。現実的には、帝国主義の植民地支配以外にあまり思いつかず、そうなると、こんな空想は忽ち挫折せざるを得ないが、何かの歴史の事情で、日本語の通じる外国があったならば、恐らく「日本が嫌になった」と移住する人もいることだろう。

アメリカで、気に入らない大統領の時代には、カナダに移住していたという人の話を耳にすることがある。

今後、機械翻訳が更に発展すれば、言葉の壁は、ある程度、克服されるだろう。どこにでも移動できるとなれば、住みやすい場所を求めて生まれ故郷を去る人も増えるのではないか。それとも、メタバースとの往き来で充分だろうか？

住みにくい国からは人が出て行き、新しい人が入ってくることもない。逆も真なりで、これは、好循環にも悪循環にもなり得る話である。

Keiichiro Hirano

小説家。1975年生まれ。1999年京都大学在学中に『日蝕』により芥川賞を受賞。以後、『葬送』、『ドーン』、『かたちだけの愛』、『空白を満たしなさい』、『透明な迷宮』、『マチネの終わりに』、『ある男』、『「カッコいい」とは何か』など、数々の作品を発表。最新刊は『本心』（文藝春秋）、『小説の読み方』文庫化（PHP文芸文庫）。

特集

AI倫理とガバナンス

AI利用が広がり、AIと共存する社会が実現しようとしている。
しかし、例えば、データに含まれる「バイアス」によって
差別的な判断をしたり、あるいは事故を起こした場合の責任の所在など、
解決すべき喫緊の課題は多い。
どのようにAIと向き合うか、AI倫理とガバナンスについて考察する。

AI倫理とガバナンス 1

AIガイドライン、トラスト、ガバナンス

―最適なルール策定に向けて―

中央大学 総合政策学部 教授

実積 寿也 Toshiya Jitsuzumi

AIの社会実装を円滑に進めるには、社会からのトラストが必要である。
各国で導入が進む非拘束的な倫理ガイドラインを、
民間プレイヤー自身が現実の財・サービスに反映させることがAIトラストの醸成につながる。
経済産業省が提案しているComply and Explainモデルは、
それを実現するためのガバナンスメカニズムとして期待されるが、
その実践が自律的かつ持続可能なものとなるためには、市場のAIリテラシーを高めるとともに、
企業による情報操作・隠蔽（AI-trust-washing）を防ぐ仕組みを整備することが求められる。
加えて、AIガバナンスの実行者である従業員向けのインセンティブ設計が現状では不十分であり、
ルール遵守が過少となりかねないことも課題である。
AIに対する特別ルールを検討する際には、技術中立性を損なったり過重な規制となったりしないよう、
AIと人間の役割分担を踏まえた上で、AI導入によって新たに生み出される効果に着目すべきである。

キーワード
AI　トラスト　ガイドライン　ガバナンス

1. はじめに

技術進歩により人工知能（AI）の能力が飛躍的に向上した結果、その社会実装がさまざまな分野で急速に進められている。AIは大きなメリットをもたらす一方で、誤用・悪用された場合のリスクが大きいため、各国は相次いでAIガイドラインを制定し、「正しい社会実装」を後押しする環境を整えつつある。しかしながら、それらはほとんど非拘束的なルールであり、強制力を持たないため、民間プレイヤーのインセンティブとの整合性を確保しない限り、「正しさ」を長

期持続的に維持することは期待できない。

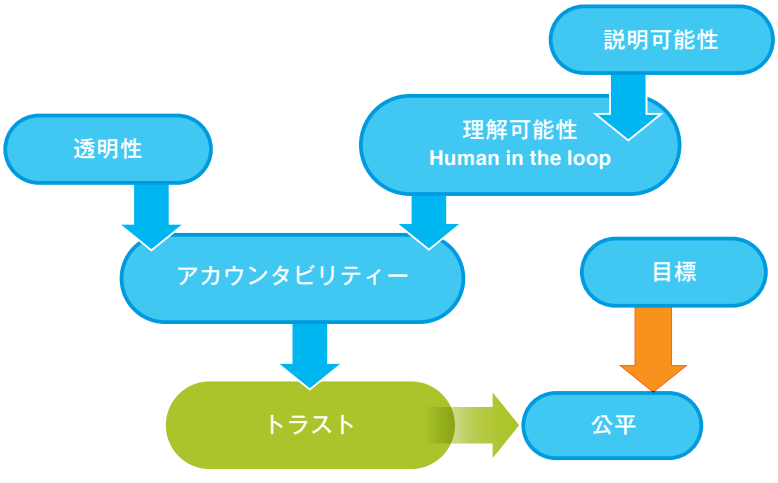
本稿では、AIガイドラインや、それを民間プレイヤーが生産プロセスに実装していくためのAIガバナンスのメカニズムについて整理するとともに、今後の検討課題の整理を行う。なお、本稿におけるAIガバナンスは、「AIを利活用したシステム・サービスを企画、開発、導入、運用するに当たって、法制度や社会規範を遵守するとともに、AIに関するリスク等を適切にマネジメントするための体制構築と仕組み作り、その実行」（AIネットワーク社会推進会議, 2021）という意味で用いる。

2. トラストとAIガイドライン

情報通信技術の急速な進歩によって社会実装が進みつつあるAIは、これまで導入されてきたさまざまな先端技術とは異なり、その倫理的側面をカバーするためのさまざまな特別ルール（いわゆるAIガイドライン）の導入が必要であると一般的に考えられている。こうした特別扱いの原因について中川（2019, 2020, 2021）は、深層学習によって自らの行動ルールを構築するAIの行動原理は人間の理解が事実上不可能であること、さらに、同じ入力に対して同じ結果が返されることが保証されていないために予測通りに機能することが常に期待できるわけではないこと、がその背景にあると指摘する。荒堀（2020）も、AIによる判断の責任所在が不明確になりやすいこと、さらに、間違いや暴走の可能性もあることを背景として掲げる。ブラックボックス化したAIを円滑に導入するためには、AIが下す判断の公平さを社会が信頼できること、すなわち、社会的なトラスト（信頼）の醸成が前提条件として必要である[1)]。AIトラストが確立して初めて、AIを利活用した財・サービスは円滑に市場展開していくことができる。

AIトラストを確立するためには、XAI（Explainable AI）といった、内田他（2021）が列挙する各種の技術的対応の導入とともに、可能である限り多数のプレイヤーによる協調行動を促す指針としての特別ルールが要請される。図表1に示すように、トラストを獲得するためには、AIという機能に対する透明性や説明可能性（および、それを基にした理解可能性）を保証し、アカウンタビリティーを確立することが要請されると中川（2019）は主張する。トラストの対象となる「公平さ」についてはユニバーサルな基準というものは存在せず、「どの要素を平等化したいのか、どういう社会を望むのか、などの外部から与えられた目標に依存して定義される」（中川, 2021, p.e36）と指摘している。これは、AIというものに対する社会的なトラストが地域や時代、さらには想定するユースケースによって異なる可能性を示唆する。

図表1　AIへのトラスト獲得に関する中川（2019）のフレームワーク



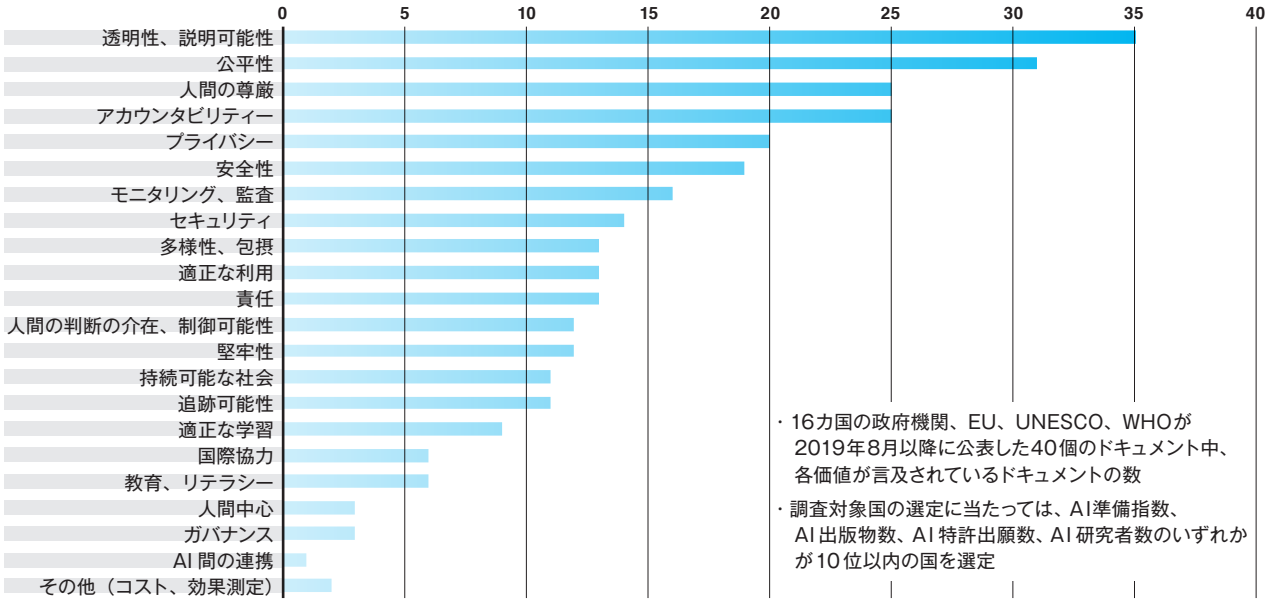
出典：中川（2019, 第5章）の記述に基づき筆者作成

AIガイドラインの採択は、OECDやUNESCOを含む世界各国で進んでおり、2020年までに採択された主なルールはRAIWG／GPAI[2]の初年度活動報告書（The Future Society, 2020）に要約されている。総務省AIネットワーク社会推進会議事務局の分析によれば、2019年8月以降にAI主要国16カ国[3]等で公表されたルールは、透明性や説明可能性、公平性、人間の尊厳、アカウンタビリティーといった要素を特に重視している（図表2）[4]。

それらルールは、わが国をはじめとして非拘束的なものとして導入されている。非拘束的なルールを企業が尊重する理由は、「AI倫理を意識した設計は、サービスや製品の価値を向上させ、企業の競争力につながるという効果もある」（AIビジネス推進コンソーシアム, 2021, p.8）と企業が認識していること、あるいは、「AIそのものの技術的な重要性や今後の活用範囲拡大の余地の大きさ、そしていったん問題が生じた場合の影響の広汎さ（中略）という特質」（小松, 2021, p.13）が重視されていること、などが挙げられている。顧客からの要望や、マスコミからの圧力などが動機付けになるという指摘もある[5]。他方で、The Future Society and Ernst＆Young（2020）が世界55カ国を対象に実施した調査では、多くの企業は、AIが引き起こす可能性のある倫理的問題よりも、実際に規制対象となっている問題に対して注目しているという結果が得られており、非拘束的ルールの実効性については、なお議論の余地が大きい。

欧州委員会は異なるアプローチを検討している。2021年4月に公表されたAI規則案[6]は、予想されるリスクに応じてAI活用システムを4分類し、それぞれに異なる規制を設定し、違反者には高額の制裁金を科す。目指すべきAI社会の姿が明確であるならば、強制力を持つルールを設定することで、民間プレイヤーの努力を結集することが期待できる。ただ、急速な技術進歩を考えると、明文規制が陳腐化するペースは速い。陳腐化への対処としてソフトローを活用すれ

図表2　AI主要国等において重視されている価値



出典：AIネットワーク社会推進会議事務局資料「資料4　AI開発ガイドライン及びAI利活用ガイドラインに関するレビュー」（2022/2/8）より筆者作成

ば、大きな不確実性の源になる。いずれにせよ、一般データ保護規則（GDPR）の例で明らかになった通り、EU規制は国際的なルールメーキングを大きく左右するため、今後の議論の動向には注視が必要である。

3. 社内AIガイドラインとAIガバナンス

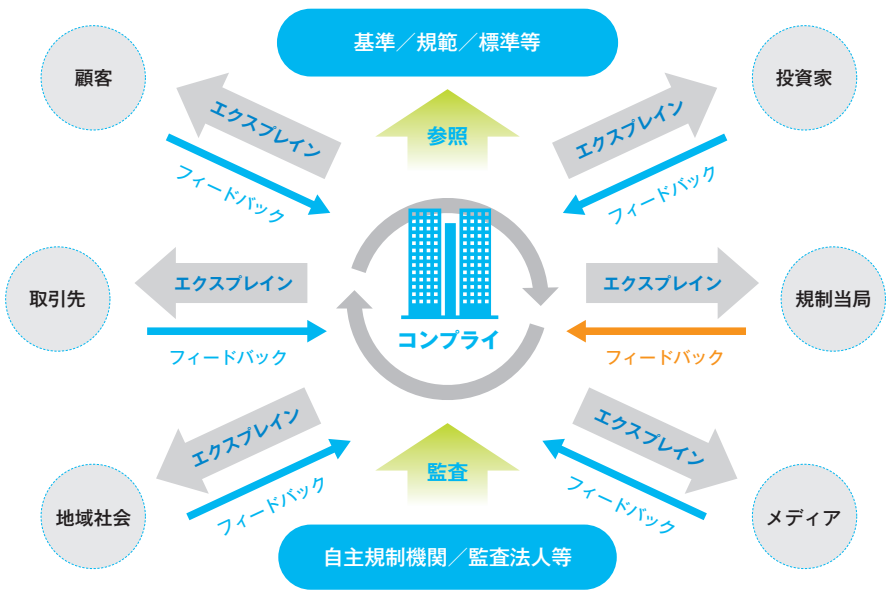
AIの利活用を行う多くの企業は、国際機関や政府での議論を踏まえ、社内向けAIガイドラインを設定している。AIサービスの競争力や、より精密なリスク管理を実現し、AI利活用企業としての高いレピュテーションを得るためには、それら社内ガイドラインの実効性を担保し（AIガバナンス）、遵守状況を市場に訴求することが重要となる。

社内ガイドラインのような自主規制には、①そもそもルール化されない「形成の失敗」、②内容が偏っているという「内容の非公正性」、③エンフォースメント・メカニズムの不在という「実効性の欠如」、④規制が市場からの認知・理解を得ていないという「公衆の認識の欠如」、⑤ルールの形成プロセスが恣意的という「民主的正統性の欠如」、さらには、⑥「国際的な非整合性の可能性」という問題がある（生貝, 2011）。多くの社内ガイドラインが内閣府のそれを参考に形成され、さらに政府の取り組みはOECDやG20、UNESCOでの議論と整合性を保っている現状に鑑みれば、わが国企業のガイドラインが直面するのは、上記③、④および⑤の問題点である。

経済産業省（2020）が提案するComply and Explainモデル（以降、CEモデルと呼ぶ）をガバナンス手段として採用することは一つの解決策となり得る。これは、企業自身が「自らのコンプライアンスの方法や考え方を設計し、規制当局あるいは市場に対して開示・説明を行い、それぞれから適時にフィードバックを受け、持続的に進化するという協創モデル」（p.42）である（図表3）。実効性を確保できる共同規制[7]を目指す場合は、例えば、新保（2020）が提案するように、情報開示を法定事項とし、実態と開示情報に乖離がある場合には事後的介入を行うという米国連邦取引委員会型の枠組みの導入が同時に要請される。ただし、共同規制を目指さない場合でも、CEモデルでは、顧客を含む外部のマルチステークホルダーの評価を最終的な判断のよりどころとするため、収益性へのインパクトを通じて、生貝（2011）が指摘する問題点③の解消がある程度は可能となる。さらに、市場からの動的なフィードバックの存在により、問題点④および⑤の改善も期待できる。

小塚（2021）は、AIルールを

図表3　Comply and Explainモデルのイメージ



出典：経済産業省（2020, 図5.2.1）に基づいて作成

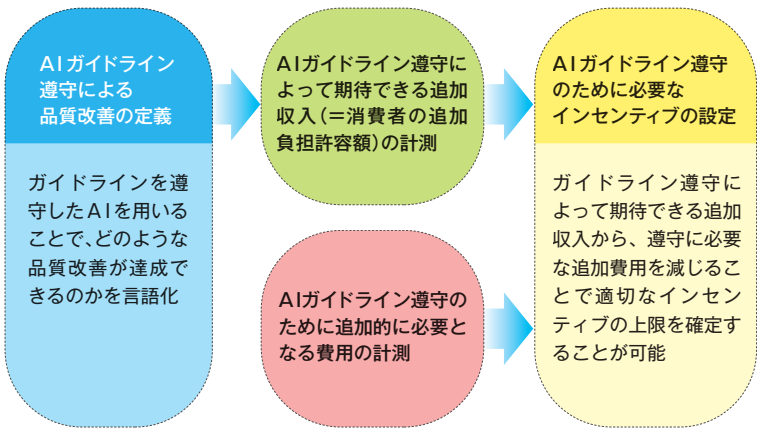
巡る議論は、何をルールとするのかという段階から、どのようにルールを実践するのかという段階に移行しつつあるとした上で、「事業者が株式会社である場合、株主共同の利益としての企業価値を最大化することがコーポレートガバナンスの理念とされる中で、コストをかけてAI原則を実施することはどのように正当化されるのか」(p.27)という民間プレイヤーのインセンティブ両立性に係る問題を指摘する[8]。AIガイドラインの遵守にはコストが必要であるため、相応の収入増がない場合は、収益性によって正当化されず、持続可能な結果が期待できない。求められるのは、高いAI倫理を求める顧客や投資家からの圧力、具体的には「より高い倫理水準を実現しているAIによって提供されるサービスを、そうでないサービスよりも高く評価する」という選好の表明である。そのため、市場への十分な情報開示とともに、AIに対するリテラシーを高める教育施策の展開が政府には期待される。

AIに関連する知識が企業側に厚く、政府側に薄いという情報の非対称性の下では、企業側の情報操作・隠蔽によって顧客や投資家の判断がゆがめられる可能性も無視できない。これは、地球温暖化対策の文脈でGreenwashing[9]と称される問題と同じであり、企業は、自社が社会に及ぼすコストをごまかすことで利潤最大化を図ろうとする。こうしたAIの信頼性に関連する虚偽行為(以降、AI-trust-washingと呼ぶ)を行わない方が企業の利潤最大化と整合的となるような制度設計は必須である。

AI-trust-washing抑止対策の一つが、従業員への適切なインセンティブ提示である。最適な社内ガイドラインが存在しても、実行に移されなければ意味がない。適切なインセンティブを設定するためには、①ガイドライン遵守により達成される最終生産物(サービス)の品質改善の定義、②当該品質改善分に対する利用者の追加負担許容額の計測、③ガイドライン遵守に必要な追加的費用の計測、という三つの作業を行った後、②と③の差分を上限[10]として設定する必要がある(図表4)。ただし、こうして設置されたインセンティブによっても得られる解は最適水準にはまだ届かない。ガイドラインを遵守することで得られるバリューチェーンを通じた経済的波及効果は、こうした手続きでは考慮できないためである。

わが国のレジ袋有料化に見られるように、金銭インセンティブによって「正しい」行動が促進されたという事例ばかりではない。金銭インセンティブが設定されたことでルールの遵守度が低下したという事例(Gneezy and Rustichini, 2000)が存在する点には注意を払うべきである。そのため、インセンティブは金銭的なものである必要はない。例えば、人工知能学会が2017年2月に定めた倫理指針は強制力を持たず、非遵守への罰則は存在しないが、同指針を満たした成果のうち優れたものに対しては表彰(「AI ELSI賞」)が設定されており、指針遵守に対する非金銭的なインセンティブとなっている。

図表4　社内AIガイドライン遵守のためのインセンティブ設定プロセス



4. AI 規制の範囲

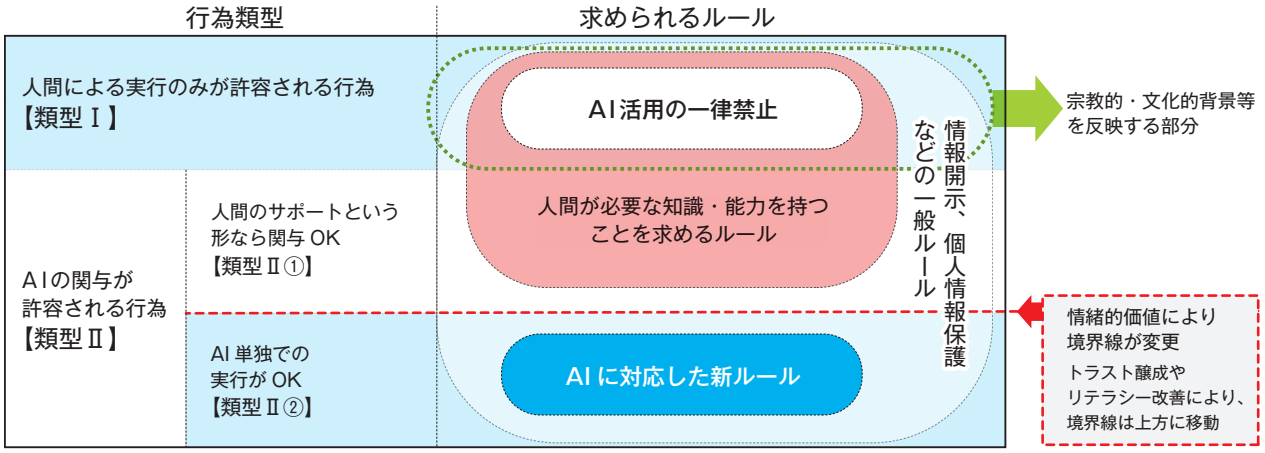
OECD の「AI に関する理事会勧告」に列挙される5原則や、内閣府の「人間中心の AI 社会原則」に含まれる7原則の全てが、AI の導入によって新たに必要になったものではない。福岡(2020)は、既存の AI 倫理原則には、(ⅰ)AI に特有なもの、(ⅱ)科学技術一般に対して従来から要請されてきたもので AI の登場によっても変化がないもの、(ⅲ)従来から要請されてはいたが、AI の登場によってその重要性を増したもの、が混在していると指摘する。

総務省や内閣府における議論では、これらのカテゴリーを区別することなく議論が進められたが、(ⅱ)や(ⅲ)を対象に「AI であること」を理由とする加重規制を求めることは、技術中立性を損ない、効率的なイノベーションを妨げる。同様の配慮は、民間企業が定めるガイドラインにおいても必要である。AI によって可能になった機能やサービスに着目して制度設計を行わない場合は、社内資源の最適配分が損なわれ、市場競争力や収益性に悪影響が生じる。他方、(ⅰ)についても、one size fits all 的なルールは存在しない。社会と AI の関係性については、宗教的・文化的背景等の影響を受けて日本と欧米では大きく異なるため(Gal, 2021 など)、欧州委員会などの枠組み[11]をわが国にそのまま導入することは適当ではない。

われわれの日常行為は、図表5に示す通り、類型Ⅰ「人間による実行のみが許容される行為」と、類型Ⅱ「AI の関与が許容される行為」に二分することができる。さらに後者は、類型Ⅱ①「人間が最終的に意思決定を行うという条件で AI の関与が許容される行為(＝人間のサポートという形であれば AI の関与が許容される行為)」と、類型Ⅱ②「AI 単独での実行が許容される行為」に分類できる。仮に欧州委員会の AI 規則案を同じフレームワークで論じるとすれば、許容できないリスクをもたらすために一律禁止が求められる AI サービス(第5条1項)は類型Ⅰに、それ以外のものは類型Ⅱに対応する。

類型Ⅰは当該国の宗教的・文化的背景等を反映する部分であり、この類型の行為に対しては AI 活用の全面禁止という措置が導入され、必要に応じて実行者である人間の知識・能力を確認するルールが導入さ

図表5　AIに対応した新ルールが必要な領域



れる。例えば、わが国では現時点でAIが運転免許を得ることは認められていないし、人間であったとしても一定年齢以上で、かつ一定の知識や運転技術が求められる。類型Ⅱ①でのAIの役割は、これまでの非AI機器が実現するものと本質的には同値であるため、「AIであること」を理由とする特別ルールは技術中立性を損なう。この類型に求められる「利用者の合理的な選択を実現するための情報開示」や、「機械学習のプロセスにおいて重要視される個人情報保護」については、必ずしもAIを原因とするものではない。ただし、類型Ⅰと同じく、人間の知識・能力の確認、特にこの場合はAIを道具として活用できる十分な知識・能力を持つことを確保するルールが求められる場合がある。それに対し、類型Ⅱ②は、AIによって初めて出現した行為群であり、情報開示や個人情報保護といった類型横断的なルールに加えて、AIが当該行為を行い得る能力を持っているか否かの確認や、事故発生時のアカウンタビリティー確保など、AIに対応した新たなルールの導入が求められる。第2節で論じたAIトラストの醸成が主要な対象とするのもこの類型である。なお、AIが十分に発展し、能力的には人間の代替が可能になった場合、類型Ⅱ①と類型Ⅱ②の境界は、当該国の消費者が「人間の温かみ」といった情緒的価値にどの程度重きを置くかによって左右される。社会的なトラストの醸成や、市場参加者のAIリテラシーの改善が進めば、将来的には類型Ⅱ①が縮小することも見込まれる。

このように整理すると、各プレイヤーが新たに整備しなくてはならないルールやガバナンスメカニズムはそれほど広範なものではなく、さらに、その範囲についてはAIを巡る社会情勢や消費者感情の変化に応じて適宜見直す必要があることが明白になる。

もちろん、実際のルール設計に当たっては、図表5の分類はさらに精緻化が必要である。類型Ⅱ①と類型Ⅱ②の境界をどこに設定するのか、各類型にどういったルールを求めるのかについては、各AIシステムの特徴を踏まえた慎重な検討が必要であり、OECD(2022)が提案したAIシステム分類方法を活用することも一案である。

5. おわりに

汎用AIの技術的可能性が見通せない現状では、AIは人類に与えられた新しい道具にすぎず、それ以上のものではない。道具であれば、あくまでも主導権は使い手側にあり、適切な使い方を考えることができる。道具のメリットとデメリットを冷静に比較較量して、最適なルールをデザインし、適切にガバナンスを行うことは、AIならぬ人間の責任である。AIを過度に警戒するあまりに、19世紀の英国で導入された赤旗法のようなルールを導入する余裕は日本社会としてもはやない。

Toshiya Jitsuzumi

実積 寿也

中央大学 総合政策学部 教授

東京大学法学部卒業、ニューヨーク大学経営大学院修了、早稲田大学大学院国際情報通信研究科博士課程修了。MBA(finance)、博士(国際情報通信学)。

郵政省(現 総務省)、コロンビア大学客員研究員、九州大学大学院経済学研究院教授などを経て、2017年より現職。総務省情報通信政策研究所特別研究員、国際大学GLOCOM上席客員研究員、Global Partnership on Artificial Intelligence専門委員などを務める。令和2年度情報通信月間推進協議会情報通信功績賞受賞。

注

1) この考え方は、欧州委員会が2019年のリポート（EC, 2019）で提唱した「信頼あるAI（Trustworthy AI）」とも共通点が多い。そこでは、「法律に従うこと」、「倫理的であること」に並ぶ三つ目の要件として、意図しない害を及ぼさないという「システムとしての堅牢さ（robustness）」が要求され、それを達成することで社会の側にはトラストが生まれる。

2) 「人間中心」の考えに基づく責任あるAIの開発と使用に取り組む国際的なイニシアチブとして2020年6月に設立されたGlobal Partnership on AI（GPAI）においてAIルールやガバナンスをメインテーマとして取り扱うResponsible AI Working Group（正式名称はResponsible Development,Use and Governance of AI Working Group）。筆者は活動開始時より参加。

3) アメリカ合衆国、カナダ、英国、イタリア、オランダ、スウェーデン、デンマーク、ドイツ、ノルウェー、フィンランド、フランス、インド、韓国、シンガポール、中国、オーストラリア。

4) 各国政府が重視する価値が一定のコンセンサスを持っていることは、The Future Society and Ernst & Young（2020）が2019年末から2020年にかけて55カ国を対象に行ったアンケート調査でも確認されている。

5) 2022年3月1日に開催されたAIネットワーク社会フォーラムにおけるFrancesca Rossi氏（IBMフェロー、IBM AI倫理グローバルリーダー）の発言。

6) Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence（Artificial Intelligence Act）and Amending Certain Union Legislative Acts, COM/2021/206 final

7) 欧州議会等が関わる機関協定（European Parliament, Council, and Commission, 2003）では、「立法機関によって定義された目的の達成を、その分野で活動する主体（経済的主体や社会的パートナー、NGOや共同体などを含む）に委ねる法的措置のメカニズム」（生貝, 2011, p.23）と定義されている。

8) 同様の指摘は、舟山（2020）、実積（2021）にもある。

9) Greenwashingは、Oxford English Dictionaryによれば「(a) 個人、会社、製品などを環境に配慮していると偽って、（一般の人を）誤解させたり、（世間やメディアの注目に）対処すること。(b)（企業やその事業などを）環境に配慮していると誤認させること」（日本語訳は筆者）と定義されている。主なものは、TerraChoice（2010）が七つに類型化し、関連研究はde Freitas Netto et al.（2020）が整理している。

10) 生産物市場および生産要素市場が完全競争条件を満たすのであれば、差分がインセンティブの適切水準と等しくなる。また、差分が負の場合は、当該ガイドラインを遵守することが企業にとって損失であり、長期持続可能ではないことを意味する。

11) 欧州委員会提案では、内包するリスクの大きさによって規制の濃淡を設定している。ただし、提案前文冒頭の第1パラグラフでは「EUの価値観に適合した人工知能の開発、マーケティング、利用について、統一された法的枠組みを定める」（日本語訳は筆者）ことが目的であると明記されており、あくまでもEUの価値観実現が目的である。

参考文献

AIビジネス推進コンソーシアム（2021）「企業活動にAI倫理を導入していく上での注意点と提言～リスクベース・アプローチを踏まえた検討～」 https://aibpc.org/?p=1523

AIネットワーク社会推進会議（2021）「AIガバナンスに関する取組事例」AIネットワーク社会推進会議メール回覧資料 https://www.soumu.go.jp/main_content/000770820.pdf

荒堀淳一（2020）「AIの責任と倫理（第3回） AI倫理に対する企業の取組み（2）」『NBL』1172, 67–71.

生貝直人（2011）『情報社会と共同規制 インターネット政策の国際比較制度研究』勁草書房.

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

参考文献

舟山聡 (2020)「AIの責任と倫理 (第2回)　AI倫理に対する企業の取組み (1)」『NBL』1170, 71–76.

実積寿也 (2021)「AIルールを巡る議論の変質」『日立総研』16 (2),18–21.

経済産業省 (2020)「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」　https://www.meti.go.jp/press/2020/07/20200713001/20200713001-1.pdf

小松岳志 (2021)「『AIとガバナンス』の企業における実践論―企業経営者にとっての『AIとガバナンス』の重要性―」『旬刊商事法務』2277, 13–18.

小塚荘一郎 (2021)「AI原則の事業者による実施とコーポレートガバナンス」『情報通信政策研究』4 (2), 25–43.

中川裕志 (2019)『裏側から視るAI　脅威・歴史・倫理』近代科学社.

中川裕志 (2020)「AI倫理指針における課題」『人工知能』35 (6), 845–854.

中川裕志 (2021)「6.　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』62 (6), e34–e39.

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

内田尚和, 鍛忠司, Blake, N., 間瀬正啓, 大橋洋輝, Ghosh, D., Gupta, C., 直野健, 高田実佳 (2021)「AIのトラストとガバナンスを支える研究開発」『日立評論』　特別増刊号, 20–27.

European Commission, Directorate-General for Communications Networks, Content and Technology (2019) Ethics guidelines for trustworthy AI. https://data.europa.eu/doi/10.2759/177365

European Parliament, Council, and Commission (2003) Interinstitutional agreement on better law-making. (OJ C 321, 31.12.2003, p. 1–5.). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003Q1231 (01)

de Freitas Netto, S.V., Falcão Sobral, M.F., Bezerra Ribeiro, A.R., and da Luz Soares, G.R. (2020) Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32 (19), 1–12. https://enveurope.springeropen.com/track/pdf/10.1186/s12302-020-0300-3.pdf

Gal, D. (2021)「AIガバナンスに関する多国間主義における日本の役割」『人工知能』36 (2), 181–186.

Gneezy, U. and Rustichini, A. (2000) Pay enough or don't pay at all. *Quarterly Journal of Economics*, 115 (2), 791–810.

OECD (2022) OECD framework for the classification of AI systems. OECD Digital Economy Papers, No.323.

TerraChoice (2010) The sins of greenwashing: Home and family edition. https://www.twosides.info/wp-content/uploads/2018/05/Terrachoice_The_Sins_of_Greenwashing_-_Home_and_Family_Edition_2010.pdf

The Future Society (2020) Areas for future action in the responsible AI ecosystem. https://gpai.ai/projects/responsible-ai/areas-for-future-action-in-responsible-ai.pdf

The Future Society and Ernst & Young (2020) Bridging AI's trust gaps: Aligning policymakers and companies. https://thefuturesociety.org/2020/07/22/report-launch-bridging-ais-trust-gaps-aligning-policymakers-and-companies/

AI倫理とガバナンス 2

人工知能とリスク分析文化

名古屋大学 大学院 情報学研究科 准教授

久木田 水生 Minao Kukita

現在、人工知能は社会のさまざまな場面で使われており、
産業や経済を活性化することが期待されている一方で、さまざまな問題も懸念されている。
本稿では人工知能の問題を近代から現代に至る歴史的文脈に位置付けて、
広い視野から考察することを試みる。
具体的に言うと、本稿では現代の人工知能の隆盛が単なる経済的な合理性によって
引き起こされているのではなく、近代以降に社会に浸透してきた、
人間のあらゆる側面を定量的に測定しようという強い欲求に関係していると論じる。
そしてその観点から人工知能と社会の将来についての展望を示す。

キーワード
二つの文化　科学的管理法　測定執着

1. はじめに

人工知能は現在、最も急速に発展しているテクノロジーの一つであり、経済と産業の起爆剤になることが期待されている。その一方で、それは社会の構造、人間の生活、人間同士の関係、さらには私たちの思考さえも大きく変化させるポテンシャルを持っている。そしてその変化は、必ずしも明るいものばかりではない。特に最近では人工知能を利用したプロファイリングが特定の人種やジェンダー、社会的集団の人々にとって不利なバイアスを持つこと、IT 企業や政府による過剰な監視とデータの乱用を助長していること、人々の思考や行動に介入していることなどが問題視されている。それ故に現在、人工知能の危険性について、あるいは人工知能に関する倫理的指針や法による規制について盛んに議論されている[cf. [2] [3] [9] [12] [15] [18] [19]]。本稿ではこういった問題を近代から現代に至る歴史的文脈から考察したいと思う。具体的に言うと、本稿では現代の人工知能の隆盛が単なる経済的な合理性

によって引き起こされているのではなく、近代以降に社会に浸透してきた、人間のあらゆる側面を定量的に測定しようという、時として非合理的なまでに強い欲求と関係していると論じる。そしてその観点から人工知能と社会の将来についての展望を示そうと思う。

2. 近代と二つの文化

イギリスの物理化学者であり作家でもあったC・P・スノーは1959年にケンブリッジ大学で「二つの文化と科学革命」と題された講演を行った[13]。ここで「二つの文化」とは、自然科学と人文学のことを指している。スノーは現代の社会で、この二つの知的文化の領域が分断されており、異なる領域で教育を受けた人々の間ではお互いを理解することが困難になっている状況に警鐘を鳴らした。

スノーの記述からは、伝統的な知的文化（すなわち人文学）に属する人々が、世界を大きく変容させている新しい知的文化（すなわち自然科学）の価値を認めようとせず、そして自分たちの文化の影響力が衰えていく現実から目をそらしている様子がうかがえる。ヨーロッパの近代を特徴付ける仕方はさまざまあるだろうが、一つの見方として、自然科学の台頭によって人文学の地位が相対的に低下していった時代ということができるだろう。

もちろん近代以前にも自然科学はあった。しかし近代においてそれは以前とは比較にならないほど大きな影響力を持つようになっていった。歴史学者のユヴァル・ノア・ハラリ[5]は近代の科学の特徴として、(1)無知の自覚、(2)観察データと数学の使用、(3)新しい能力の獲得という3点を挙げる。ここでは(2)に注目しよう。近代科学は観察データにフィットする数学的理論を構築するという方法を確立した。この方法によって、ひとたび理論が確立されたならば、数式の扱いに慣れた人間であれば誰でも、単純な数式に値を代入するだけでさまざまな現象について正確な予測ができるようになった。このようにしてさまざまなシステムについて、そこから得られるデータを収集し、それにフィットする数学的モデル・法則を作り、それによって未来の現象を予測することが、科学的実践のルーティンとなった。

近代科学のこのパラダイムを確立したのはニュートンである。ニュートンの力学理論は地上の物体から天体までもカバーする高い一般性を持つが、それでもそれで扱えるシステムは比較的単純なものに限られていた。例えば単純な力学法則に従う物体であっても、三体以上の物体が相互に影響を与えているような場合は、その振る舞いを運動方程式の枠組みで定式化することには限界があった。まして無数の分子が相互作用して生じる流体の運動や気象現象など、あるいは人間や社会システムの振る舞いなどは正確な記述や予測、定量的測定が困難であった。このような複雑なシステムの振る舞いの予測を可能にしたのが統計学である。統計学では複雑なシステムを構成している個々の要素に注目するのではなく、システム全体の趨勢（すうせい）に焦点を当てる。そうすることで、システムの個々の構成要素については、どのくらいの確率でどのように振る舞うかを予測することができる。統計学によって、人間の集団もまた数学理論によって、定量的に扱われることになった。

自然科学におけるこのような動きは、人間の個人性を重んじる人文学の風潮と対照的である。ルネサンス

的な人文主義の精神の一つの特徴に、人間の持つ無限の可能性に対する信奉があった。ルネサンス期の人文主義者のピコ・デッラ・ミランドラは「人間はその生活を自分で選択しているのであり、宇宙のヒエラルキーのなかで固定した位置を占めつづける必要などなく、カメレオンの如く変幻自在だ」([6] p. 232)と述べた。同じくルネサンスの人文主義者であるレオン・バッティスタ・アルベルティは「人間は意志すれば、あらゆることができる」([17] p. 18)と述べた。ルネサンス期に続いて興った啓蒙主義の運動においては、全ての人間が他者によって奪われることのない尊厳、自律性、権利を持つと主張された。このような思想的潮流の中で、それまで封建的権力や宗教的権威によって縛られていた個人が、自由を獲得して、より流動的に活動するようになった。

とはいえ個人は完全に自由になったわけではない。ルネサンス研究者の池上俊一[6]は、ルネサンス期において伝統的な権威から解放された個人は、今度は家族や都市という新しいイデオロギーの中に取り込まれていったと指摘している。そして近代になると、人々はさらに新しい強大なシステムである国家や経済というイデオロギーの中に取り込まれていくことになる。こうして人間が自由を得て流動的になったことは皮肉な結果をもたらした。強大な経済的システムの中に取り込まれた個人は、いくらでも取り換えの利く部品になったのである[1]。その際、経済システムは、産業革命以来それらに大きな力を与えてきた科学と数学を大いに活用した。人間は数字によって動態を把握され、数字によってその経済的有用性を測定された。

科学史学者の隠岐さや香は人文社会科学と自然科学について、次のような印象的な特徴付けを行っている。

> 諸学問の中には、「人間」をバイアスの源として捉える傾向と、「人間」を価値の源泉と捉える傾向とが併存しています。前者の視点は、自然科学の営みにおいて特に成熟しました。後者は、人文社会科学に深くかかわっています。([11] p. 246)

いささか乱暴にまとめるならば、人間の個性、その変幻自在な発展の可能性を称揚した人文学の文化が衰退し、それに代わって人間的なバイアスを取り除き、さまざまな事象を数学的・科学的に扱う文化が発展したのが近代という時代だった。

3. 測定のプレッシャー

近代のこのような風潮を象徴するのが、1911年にアメリカの工学者、フレデリック・ウィンズロウ・テイラーが提唱した「科学的管理法」である。これは工場での生産工程を細かい要素に分解した上で、それぞれの作業におけるアウトプットの基準を制定し、それに基づいて労働者のパフォーマンスを測定するというものである。そして労働者は測定されたパフォーマンスに応じて報酬を支払われる。

科学的管理法においてはあらゆる仕事が細かいルーティンに落とし込まれ、労働者には早く正確にルーティンをこなすことが求められるようになる。これによって画一化された製品がどこででも効率的に作れるようになった。科学的管理法は製造業にとどまらず、他の多くの分野へと広がっていった。

科学的管理法は確かにある分野では生産性を向上させるための有効な手段になり得る。しかしながらこの方法が常により良い結果を生み出すと考えるのは誤り

である。歴史学者のジェリー・Z・ミュラーは、医療・教育・警察・軍事などの分野では、定量的なパフォーマンス評価に基づく管理方法が導入されたことで、かえって悪い結果が引き起こされている、と論じている[10]。そのような失敗は得てして、定量化するのが難しい事柄を無理矢理に定量化しようとすることによって引き起こされる。

例えば私の所属している学問の世界でも定量的評価のプレッシャーが年々大きくなっている。大学は文部科学省からさまざまな指標に基づいた評価を下され、研究者は大学からさまざまな指標に基づいた評価を下されている。その際に参照されるのは、発表した論文や書籍の数、論文が引用された数、担当した講義の数、講義を受けた学生によるアンケートの結果などなどである。しかし実際に書いた論文や書籍の内容、授業の内容が評価の俎上（そじょう）に載せられることはない。それらは定量化できないからである。おそらくほとんどの場合、教員の評価をする側の人間は各教員の論文や書籍の中身をちらっと見ることもないし、講義の内容を調べることもない。現在の研究者たちは、その最も重要な仕事である論文や講義の中身を知らない人間によって評価されている。

このことの弊害はさまざまである。第一に、数値化できない部分は顧みられなくなり、数値化されるところばかりに注意が向かう。例えば講義の評価が学生アンケートの結果に基づくのであれば、学生に受けが良いような授業をすればよいということになる。もう一つの弊害は、評価に関連する業務に無駄な労力が費やされているということである。今日の組織の管理部門は意味のない、むしろ悪影響すらあると分かっている評価のために多くの人員と予算を投入している。管理される側も評価のために管理部門に要求されるデータや書類を作ることに忙殺され、また評価をどのようにして向上させるかということに腐心する（人類学者のデヴィッド・グレーバー[4]はこのような、本質的に必要なく、そしてやっている本人たちでさえその意義を感じていないような仕事を、「ブルシット・ジョブ」と呼んでいる）。

上述のジェリー・Z・ミュラーは、現代の社会においては、組織は人々の行動を監視し、パフォーマンスを定量的に評価し、それに応じた賞罰を与えるべきだ、という思い込み、「測定執着」が蔓延（まんえん）している、と説く。測定執着は、一つには多くの人が抱いている「投資にはそれに見合うリターンがなければならない」という信念に根差している。そのために組織は出資者に対して、出資に見合う価値を創出しているかどうかを説明する責任があると考えられる。それはビジネスの世界では当然のことかもしれない。しかし近年ではこの考えが医療や教育、あるいは行政などの公共セクター、果ては個人の私的生活の領域にまで侵食している。行政など公共セクターの仕事は、出資に見合う成果がはっきり表れるから行われるものではない。医療や教育についても、同様のことがいえる。それらはビジネスの観点だけからは捉えられない側面を持つ。にもかかわらず現代の社会では、ありとあらゆる領域において、可能な限りビジネス的な観点から論じるべきであるという規範が浸透しており、それが測定執着に拍車をかけているように見える。

4. 人工知能とリスク分析文化

「人工知能」という言葉は、多様なテクノロジーに対

する曖昧な総称である。本稿ではビッグデータに基づいて機械学習を行い、人間を分類したり評価したりすることに使われる人工知能に話を限定する。以下では「人工知能」という言葉によってそのようなシステムを指すものとする。

現代の人工知能がやっていることは、与えられた大量のデータから共通するパターンを見つけ出し、そのパターンに基づいて（確率的な）分類や予測を行うことである。例えば大量の猫画像から猫に共通するパターンを抽出し、新しい画像についてそれが猫であるかどうかを識別できるようになるという具合である。

これだけだとそのインパクトは分かりづらいが、機械学習の真価は実世界のビッグデータと組み合わされたときに発揮される。今日、私たちの多くの行動のデータがさまざまなデバイスを通じて収集・保存されている。そして人工知能はこの膨大なデータ（ビッグデータ）に基づいて、人々を分類し、人々の振る舞いや属性、性格や能力を予測、推測する。データが集まれば集まるほど予測できることが増え、また予測の精度も増していく。

ビジネスにおいては人々の行動を正確に予測できることは大きなアドバンテージである。いつ、どこで、どういった需要が、どれくらいの確率で発生するかを正確に知ることができれば、企業はその分だけ無駄なく効果的なアクションを起こすことができ、より多くの利益を上げることができる。さらには適切なタイミングで適切な情報を与えれば、人々の行動を誘導することもできる。何億というユーザーを抱える企業であれば、予測の精度がほんのわずかでも上がれば、それによって得られる利益も莫大（ばくだい）なものになる。だからこそAmazonやGoogleなどの巨大IT企業はますます貪欲にデータを収集し、そして人工知能の開発に巨額の投資を行っている。人工知能はこういった大金持ちが超大金持ちになることに、現在最も有効に活用されている。社会学者のショシャナ・ズボフはこれを「監視資本主義」と呼ぶ[20]。

しかし人工知能の活用に関心を持つのは、そういった巨大IT企業ばかりではない。人工知能は利用者の利益関心に沿って、他者についてさまざまな予測・推測を行い得る。典型的な例を挙げれば、労働者としての能力や適性、犯罪を実行する可能性、ローンを返済する能力、特定の病気にかかる（あるいはかかっている）可能性、交通事故を起こす可能性、配偶者としての相性、ある商品にどれくらいの金額を支払いそうか、ある政党を支持しているかどうか、などなどを予測・推測するために人工知能が使われている。

意思決定を行う際に、あり得る選択肢のそれぞれに伴う潜在的な利益と損害を見積もることを、工学や経済学の分野では「確率論的リスク分析」あるいは単に「リスク分析」と呼ぶ。リスク分析においては、ある選択肢を採ったときに、どのような事象がどれくらいの確率で生起するかを見積もることが必要になる。しかし従来は人間や社会のような複雑なシステムについて、特定の振る舞いの生起確率を正確に予想することは難しかった。それ故に人間に関するリスク分析はしばしば多くの仮定に基づいた不正確なものにならざるを得なかった。しかしビッグデータと人工知能は人間の振る舞いについて、従来よりはるかに正確な確率的予測を可能にした。人工知能は、人間や社会を対象にした確率論的リスク分析のための革新的なツールである。

確率論的リスク分析は企業や政府機関が意思決定を

行う際に行われるものであるが、人工知能がますます多くの場面で使われるようになると、人工知能を使ったリスク分析もまた私的領域でもカジュアルに行われるようになるだろう。人々が個人的な意思決定――どの映画を見るかとか、どこで食事をするかといった日常的なものから、進路や就職や果ては結婚相手といった人生における重要な選択にまで――人工知能を使ってリスク分析を行うかもしれない。社会学者のデイヴィッド・ライアンは、現代においては政府や企業などの組織だけではなく、個人が進んで相互に監視する「監視文化」が浸透している、と指摘する[8]。人工知能が手軽に使えるようになっていくにつれて、「リスク分析文化」とでも呼ぶべき慣習が浸透していくかもしれない。

5. 人工知能の陥穽（かんせい）

人工知能は3節で触れたテイラー主義、測定執着と相性が良い。と言うかそれは一続きのものとして考えるのが適当だろう。数学者、データサイエンティストであり、ビッグデータや人工知能についての啓発活動を行っているキャシー・オニールは、人間について予測を行ったり、評価したり、順位付けをしたりするために使うことを意図して作られた統計的・数学的モデルの中には、社会に大きな害を与えているものが存在していると指摘する。そしてそれらを「数学破壊兵器」と呼んでいる[12]。オニールの数学破壊兵器は、古典的なパフォーマンス測定法と、より新しい人工知能の両方を包含する概念である。

古典的なパフォーマンス測定と同様、人工知能も「人間と異なり、バイアスがなく、公平で、客観的かつ正確である」としばしば喧伝（けんでん）される。これは大きな間違いである。人工知能も多くの場合、特有のバイアスを持つ。人工知能のバイアスの源泉は主に二つある。一つは判断のためのアルゴリズム、もう一つは学習のもとになるデータである。ざっくり言うと、アルゴリズムには作成者の偏見や先入観や意見が、データには社会の持つそれらが反映される。

まずアルゴリズムについて。私たちは人工知能を使って何らかの属性を、その他の属性から判断する。この際、後者の属性として何を使うかを選択するのは、システムを作る人間である。そしてその選択に人間のバイアスが反映される。分かりやすい例を取り上げよう。愛知県のある保育園運営会社は保育士を採用するための面接で人工知能によって「笑顔度」を測定するというシステムを導入した[14]。このシステムは面接中の応募者の顔を判定して、笑顔でいる時間の長さを計測する。そして面接担当者の評価するコミュニケーション能力を50点満点、AIの測定する笑顔度を50点満点として最終的な評価を決定する。ここには「面接中に長い時間笑顔でいられる人間ほど良い保育士になる」という前提がある。この前提は客観的な事実ではなく、システムを作った人間の思い込みかもしれない。さらに、笑顔度が人間の面接官の評価と同等の重要性を持つというのもシステムを作った人間が決めたことである。このようにして評価アルゴリズムには作成者のバイアスが反映される。

アルゴリズムのみではなく、人工知能が学習するのに用いられるデータにもバイアスがかかり得る。例えば過去の犯罪のデータは、その社会が特定の人種に対して差別的である場合は、公平なデータではない。なぜなら社会がその人種の人間を犯罪予備軍と見なし

て、ほかの人種の人間よりも厳しく監視していたために、同じ犯罪をしてもその人種の人間は捕まり、ほかの人間は見逃されるということの結果かもしれないからである。このようなデータに基づいて予測を行い、さらにその人種の人間を厳しく監視した結果、ますますその人種の人間の検挙率が高まるということになる。実際に、アメリカで再犯の可能性などについて推測するために使われ、量刑の決定にも影響を与えているCOMPASというシステムでは黒人の再犯率を白人の再犯率よりも高く見積もると指摘されている[16]。

AC Global Risk社が開発している「遠隔リスク評価(Remote Risk Assessment;RRA)」というシステムには人工知能による人間評価のさまざまな問題点が顕著に表れている[7]。このシステムは電話越しの10分程度の会話(決められた質問に対して母国語で答える)のみによって危険人物であるどうかを判定するという。「移民を徹底的に精査せよ」というトランプ元大統領の要求に対する答えとして、AC Global Risk社はRRAを「アメリカや他の国が現在、直面している歴史的な移民危機に対する究極のソリューション」と喧伝している。AC Global Risk社はソフトウェアの詳細に関する質問に答えることを拒否しているが、公開されている材料をレビューした専門家はこれを「たわごとbullshit」、「いかさまbogus」と呼んでいる。アメリカの入国審査では、人を捜査したり入国拒否したりする際、喋り方や見た目が口実として使われている。RRAはそのようなバイアスを「ルーティンとして蔓延させ、一見『客観的』に見せるかもしれない」と専門家は危惧している。

古典的なパフォーマンス測定と同様、人工知能も使いどころを慎重に考えなければ有害な結果を引き起こす。それらは喧伝されるほど正確でもないし、バイアスを免れているわけでもない。そしてそれらはしばしばすでに困難な状況に置かれている社会的弱者、貧しい人々、差別されている人々に、より不利な判断をする。

こういった問題が次々に明らかになっているにもかかわらず、企業や政府はアルゴリズムで人を評価することに熱心である。なぜならそれは複雑な世界における複雑な問題に対する単純で効率的な「解決」であり、自分たちのバイアスを確証してくれる都合の良い道具であり、また測定執着の蔓延する世界においては説明責任を果たしたように見せかける格好の手段になるからである。それは本当の解決ではないし、本当の説明責任の放棄ですらある。しかし人々は複雑な世界の複雑な問題に、簡単な解決があると信じたいが故に、人工知能は人間と違って公平であるというフィクションを共有している。そうして人間はますます機械可読なデータと、そこから定量的に計算、推測できる属性の束として見なされる。

6. 終わりに

本稿では近代以降に発展してきた、人間は定量的に測定できる、測定するべきだと考える文化の文脈の中に位置付けて論じた。人工知能はこの文化に連なるものであり、そしてこの文化をますます強化することに寄与するだろう。そうして人間はもっぱらリスクの源泉として見られ、人工知能を使ってそのリスクを見積もり、回避することが組織の管理者の義務と見なされる。経営コンサルタントが組織の管理者にあれやこれやのパフォーマンス測定指標を売り込むように、人工

知能を売り込む業者が組織運営上のブルシット・ジョブ・マーケットで大きな一角を占めるようになる。

他方で管理される側の人々は人工知能に高く評価されることに腐心するようになるだろう。ウェブページの管理者が検索エンジン最適化（検索エンジンのランキングの上位になるようウェブページを工夫すること）を目指すように、人々は自分を評価する人工知能に対して最適化を目指すことになる。そしてそれはもともと評価・改善するべきだったものが何だったかを忘れさせる。

私は人工知能の全てが悪いとは考えていない。科学的管理法と同様に、人工知能にも使うべきところと使うべきでないところがある。例えば複数の人間による協調や社会的なスキルが要求される場面では、人工知能による評価は難しいだろう。協調が必要となる場面で、一人の人間がどれだけ重要な貢献をしているかというようなことは簡単に数値に表せるものではない。組織の中にはスタープレイヤーもいれば、スターが活躍しやすいようにサポートする縁の下の力持ち、“unsung heroes”がいる。そういう人たちの働きは、おそらく人工知能によっては評価するのが難しいだろう。誰もがスタープレイヤーを目指さなければいけないようなプレッシャーがあるようなところでは、かえって組織全体のパフォーマンスは低下するかもしれない。

最後に、人間の性格や能力は固定されたものではなく、周囲との関わりの中でダイナミックに変化するものであるということ、良い変化、成長はしばしば信頼から始まるということを忘れないようにしよう。と言うのも人は信頼されるとその信頼に応えようと努力するものだからである。測定やリスク分析は疑いをベースにしており、相手の自発的な成長のきっかけになるようなものではない。測定執着から予測執着、リスク分析執着へと高じていくとき、人々の間の信頼関係は大きなダメージを被るのではないかということを私は危惧している。

あるいは私は悲観的すぎるのかもしれない。しかしこの先、人工知能がさまざまな場面で使われていくにつれて、それが現在ただでさえ行き過ぎている測定執着を一層強化し、さらには予測執着、リスク分析執着の文化を醸成するだろうということは、歴史の流れを見れば必然的に思われる。その中で、人間をバイアスの源としてだけではなく、価値の源泉と見なす、そして人間の無限の可能性を称揚する人文科学的、人文主義的な人間観が完全に失われないようにしたい。

Minao Kukita

久木田 水生

名古屋大学 大学院 情報学研究科 准教授

2005年、京都大学大学院文学研究科で博士号（文学）を取得。2017年より現職。専門は情報の哲学、技術哲学、人文情報学など。

著書に『ロボットからの倫理学入門』（共著、名古屋大学出版会、2017年）、『人工知能と人間・社会』（共編著、勁草書房、2020年）、『学問の在り方――真理探究、学会、評価をめぐる省察』（共著、ユニオン・エー、2021年）など。翻訳書にアンディ・クラーク『生まれながらのサイボーグ』（共訳、春秋社、2015年）、ウェンデル・ウォラック&コリン・アレン『ロボットに倫理を教える』（共訳、名古屋大学出版会、2019年）、マーク・クーケルバーク『AIの倫理学』（共訳、丸善出版、2020年）などがある。

参考文献

[1] ジークムント・バウマン、『リキッド・モダニティ―液状化する社会』、森田典正訳、大月書店、2001年。

[2] 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』、秋山勝訳、草思社、2018年。

[3] 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。

[4] デヴィッド・グレーバー、『ブルシット・ジョブ―クソどうでもいい仕事の理論』(電子書籍版)、酒井隆史、芳賀達彦、森田和樹訳、岩波書店、2020年。

[5] ユヴァル・ノア・ハラリ、『サピエンス全史―文明の構造と人類の幸福』(上・下)、柴田裕之訳、河出書房新社、2016年。

[6] 池上俊一、『イタリア・ルネサンス再考―花の都とアルベルティ』、講談社学術文庫、2007年。

[7] A. Kofman, "Dangerous junk science of vocal risk assessment", The *Intercept*, November 25, 2018. https://theintercept.com/2018/11/25/voice-risk-analysis-ac-global/

[8] デイヴィッド・ライアン、『監視文化の誕生―社会に監視される時代から、ひとびとが進んで監視する時代へ』、田端暁生訳、青土社、2019年。

[9] 村上祐子、「人工知能の倫理の現在―研究開発における技術哲学・倫理の意義」、*IEICE Fundamentals Review*, 11(3), pp.155-163, 2018年。

[10] ジェリー・Z・ミュラー、『測りすぎ―なぜパフォーマンス評価は失敗するのか？』、松本裕訳、みすず書房、2019年。

[11] 隠岐さや香、『文系と理系はなぜ分かれたのか』、星海社新書、2018年。

[12] Cathy O'Neil,*Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Penguin Books, 2017. 邦訳：キャシー・オニール、『あなたを支配し、社会を破壊する、AI・ビッグデータの罠』、久保尚子訳、インターシフト、2018年。

[13] C. P. Snow, *The Two Cultures and Scientific Revolution*, Barakaldo Books 2020.

[14] 角拓哉、「保育士に大切なのは…AIが採用面接で「笑顔度」測定」、『朝日新聞デジタル』、2020年9月7日。https://www.asahi.com/articles/ASN9762C1N8YOIPE00Y.html

[15] 平和博、『悪のAI論―あなたはここまで支配されている』、朝日新書、2019年。

[16] 平和博、「見えないアルゴリズム：「再犯予測プログラム」が判決を左右する」、2017年8月8日。https://www.huffingtonpost.jp/kazuhiro-taira/clime_b_11381184.html

[17] Bard Thompson, *Humanists and Reformers: A History of the Renaissance and Reformation*, William B.Eerdmans, 1996.

[18] 山本龍彦、『おそろしいビッグデータ―超類型化AI社会のリスク』、朝日新書、2017年。

[19] 弥永真生・宍戸常寿編、『ロボット・AIと法』、有斐閣、2018年。

[20] ショシャナ・ズボフ、『監視資本主義―人類の未来を賭けた闘い』、野中香方子訳、東洋経済新報社、2021年。

特集論文

AI倫理とガバナンス 3

AIガバナンスを支える技術

株式会社 KDDI総合研究所 AI部門
黒川 茂莉 Mori Kurokawa

株式会社 KDDI総合研究所 AI部門
国立研究開発法人 産業技術総合研究所 人工知能研究センター
麻生 英樹 Hideki Asoh

深層学習等のAI技術の発展と社会への普及の進展により、
AIの利活用の際に発生し得るさまざまなリスクを管理し、
AIによって得られる効果を最大化することを目的とする
AIガバナンスの在り方が国内外で議論されるようになっている。
本稿では、AIガバナンスを支える技術として、AIのリスクを低減させ、
管理するための代表的な技術の現状と今後の発展可能性について解説する。

キーワード
AIの説明可能性　AIの公平性　AIの品質保証　AIのプライバシー保護

1. はじめに

AIの歴史は、1960年ごろにさかのぼり、幾度かの隆盛と幻滅期を経て、今日の隆盛を示している。2010年代の深層学習の発展は、約50年間の基礎技術の積み上げ、ネット上で収集された膨大なデータ、計算機能力の向上の3点が組み合わさって実現されたもので、特にネット上で収集された画像やテキストを用いた画像認識や自然言語理解のタスクにおいては、人間の能力と同等または同等以上の性能を示すに至っている[1][2]。

高度な能力を獲得したAIの利活用は、ネットビジネスを中心としたさまざまな企業で始まっており、Google検索やAmazonレコメンドなどでも成功を収めている。成功の拡がりの陰でAIの利活用に関するリスクも顕在化しており、Amazon社の人材採用における不公平な差別的取り扱いなどの具体的な諸問題も周知の事実になりつつある。

AIは多数のパラメータを含む高度に複雑なシステ

ムであり、人間がAIの動作を完全には予測できないため、通常の技術と比べてリスクを見積もることが非常に困難である。また、AIの利活用に際し、自律的に人間が関与せず動作する場合が多いことも、リスクの制御や管理を困難にしている。

このような背景を踏まえ、OECDの「AIに関する理事会勧告」、内閣府の「人間中心のAI社会原則」などのAIの開発や利活用の際に必要な価値観が取りまとめられ、その共有が図られるようになってきた。さらに、そうしたAI原則を社会で実現するためのガバナンスの在り方についての議論も始まっている。

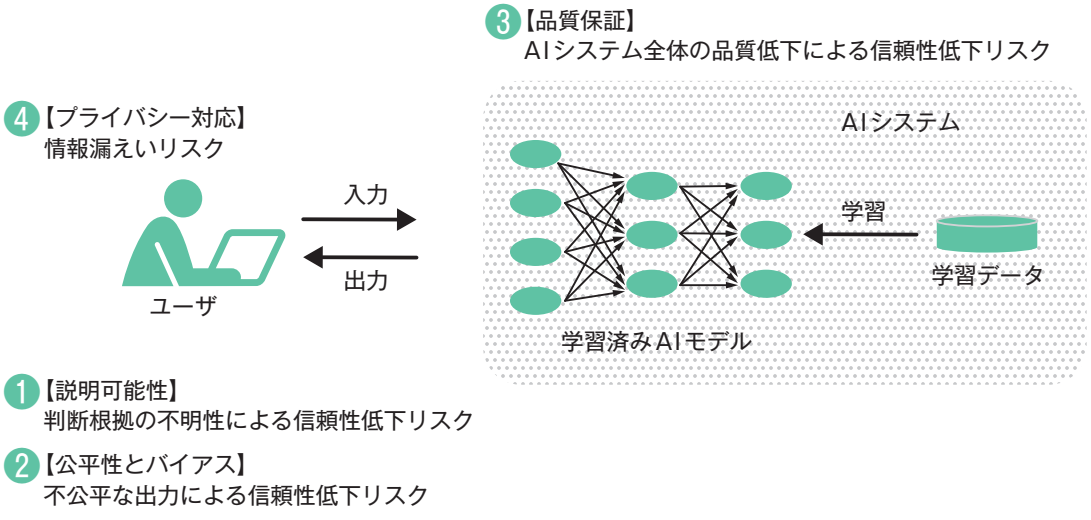
例えば、経済産業省のAI原則の実践の在り方に関する検討会は「我が国のAIガバナンスの在り方」[3]において、規制やモニタリングの望ましい姿について検討し、現時点では、法的拘束力のある横断的な規制ではなく、AIを利活用している企業を中心として幅広いステークホルダーによって議論された非拘束の中間的なガイドラインが望ましいと提言している。さらに「AI原則実践のためのガバナンス・ガイドライン」[3]では、具体的な行動目標や実践例をまとめている。

こうした社会的動向を反映して、AIの利活用の際に発生し得るさまざまなリスクを管理し、低減させるための技術の研究開発が進められている。本稿では、こうしたAIガバナンスを支える代表的なAI技術の現状と今後について紹介する。図表1に示すように、現在のAIシステムは学習データから学習したAIモデルを用いて、ユーザからの入力に対してユースケースに応じた何らかの出力を行う。そこにおいて、入出力あるいは全体の挙動に対する各種リスクが存在し、それに対応して、AIの説明可能性(2節)、AIの公平性とバイアス(3節)、AIの品質保証(4節)、AIのプライバシー保護(5節)といった技術的な課題が検討されている。以降、各課題について解説してゆく。

2. AIの説明可能性

AIの動作が説明困難となるのは、主として、深層学習モデルのような複雑なAIモデルの場合である。AIが線形回帰のような単純なモデルであれば、一定の知識があれば、回帰係数などのパラメータを読むこ

図表1　AIシステムの利活用に伴う代表的なリスクと対応技術



とによって、どの変数がAIの判断に寄与しているかを理解することができる。一方、AIが深層学習によって学習された複雑なモデルの場合は、複雑な信号伝播などの機構を持つため、動作の理解や説明が難しい。深層学習モデルの複雑さのイメージとして、図表2にいくつかの生物のシナプス数[4][5]と著名な画像認識モデル（深層学習モデル）[6]のパラメータ数を比較する。深層学習モデルは、ヒトほどではないが多数のパラメータを有する。

また、図表3のように、パラメータ数が多い複雑な深層学習モデルほど、予測精度は高い傾向が示唆されている。一方、複雑になるほど、その全てのパラメータの挙動を全てトレースすることは困難であり、トレースしたとしても動作が人間の言葉で説明できたとは言えないという状況になり、予測精度と説明力にはトレードオフがある。

そこでAIのパラメータの挙動を全てトレースすることなく、AI判断の根拠を説明してほしいという期待が生じる。図表4に代表的な方法を示す。まず、一つ目の方法として、AI判断の根拠を元データに求めるという方法がある。画像認識などの場合、例えば「猫」と判断した根拠として、元画像の判断に寄与する画像の領域を特定する方法[7]がある。これは深層学習の際に、与えられた教師信号「猫」に適合するために、元データの特定の領域に対応してパラメータが鋭敏に動くことを捉えることによって可能となる。また、判断根拠として、関連が強い学習データを選択する方法[8]もある。画像に関する応用事例として、DeepMind社が、網膜の病気診断支援において、網膜の画像データをセグメント化して根拠を提示する方法を提案している[9]。

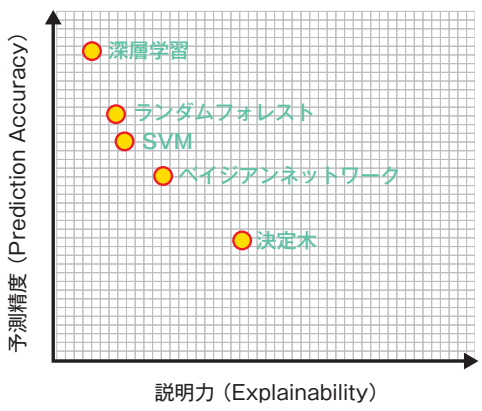
また、画像以外のデータに対しても判断根拠となる元データの特徴を捉える手法が各種提案されている。決定木（けっていぎ）のように比較的シンプルなモデルの場合、判断結果に対して元データの割合についての偏りが大きい変数を捉えることで、重要変数を定量化する方法がある。

二つ目の方法として、一般的な機械学習モデルに対しても、LIME[10]やSHAP[11]と呼ばれる手法がある。LIMEもSHAPも、複雑なAIモデルに対して、対象

図表2　生物の神経系とAIモデルの複雑さの対比

生物	シナプス数（全神経系）
ハツカネズミ	～1×10^{12}
ヒト	～1.5×10^{14}
画像認識モデル	**パラメータ数**
ResNet50	25,636,712
InceptionResNetV2	55,873,736

図表3　AIモデルの説明力と予測精度の関係



出典：DARPA-BAA-16-53: Proposers Day Slides を基に作成
（https://www.darpa.mil/program/explainable-artificial-intelligence）

データの周辺データに限定して説明のための代理モデルを作成し、代理モデルに基づいた判断根拠を提示する。LIMEでは代理モデルとして線形モデルを用いる。また、SHAPはさらに、判断結果への貢献度合いを各説明変数に分配する際に、協力ゲーム理論におけるShapley値を利用するところに特徴がある。

三つ目の方法として、深層学習の複雑なモデル全体を単純化することにより、説明しようというアプローチがある。このような手法を、アルコールなどの蒸留工程になぞらえて蒸留と呼び、複雑な深層学習モデルを教師モデルとし、その入出力結果と似た結果を示すより簡単でパラメータ数の少ない生徒モデルを生成するという方法がある[12]。生徒モデルとして可読性の高い決定木などのモデルを採用することにより、どのような入力を与えるとどのような出力が得られるかが理解しやすくなる。

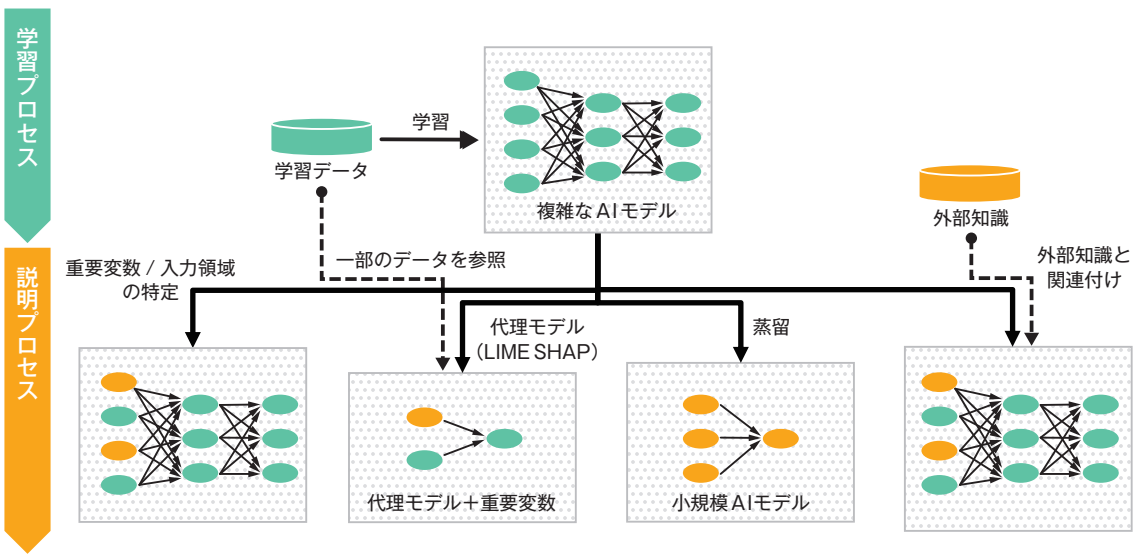
四つ目の方法として、AIの判断根拠を、人間にとって可読性の高い外部知識（知識グラフ）などに求める手法[13]も提案されている。

以上のようにAIの説明可能性に関するアプローチはさまざまあるが、AIのユーザが求める説明のレベル、詳細度はさまざまであり、いかに手法を選択し、組み合わせるかについてはケースバイケースであり、統一的な方法はまだないのが現状である。また、判断根拠として対話などによる自然な説明を実現するには、対話AIなどの技術との融合的な発展が必要と考えられる。

3. AIの公平性とバイアス

人間が偏見を持ったり、不公平な判断をしたりすることがあるように、AIがユーザに提示する判断が社会的な偏見（バイアス）を持ち、社会的な公平性（フェアネス）に欠けるものになることがある。例えば、人事評価を支援するシステムの判断が、性別という属性によって大きく変わることは一般的には望ましくない。実際に、2018年10月には、米国Amazon社の、過去の履歴書等のデータで学習したAIを用いた人材採用システムが、ソフトウエアの開発などの職種に関して、女性に不利な判断を出す可能性があることが判

図表4　AIモデルの判断根拠を提示するための代表的手法



明したため、その利用が打ち切られる、というニュースが話題になった。

こうしたバイアスが生まれる原因としては、大きく以下の三つが挙げられる：1)学習用データの収集法に偏りがある、2)学習用データが現在の社会に内在する偏りを反映する、3)学習に使ったモデルやアルゴリズムが偏りを生じさせる。このうち、1)と3)については、機械学習AIの品質保証について後述するように、学習用データの収集時や、データを用いた学習を行う際に注意をすることである程度防ぐことは可能だが、2)を最初から防ぐことは難しい。

バイアスを防ぐために、例えば性別のようなセンシティブな属性を入力情報として使わなければよいのではないか、と思われるかもしれないが、それだけでは十分ではない。例えば、趣味が料理、といった情報があれば、対象者が女性である確率が高まるため、センシティブな属性を排除しても、依然としてシステムがバイアスを持つことは起こり得る。しかし、関連のある情報を全て落とすことは、システム自体の性能を大きく低下させることになるだろう。また、平均的に見た能力に差がある場合に、システムがそれを反映してしまうことは仕方がないのではないか、と思われるかもしれないが、それは、現在の社会に内在するバイアスを是認し、増強させることにつながる可能性がある。

そこで、システムが行う判断について、社会的にセンシティブと見なされる属性が関与する場合には、データを収集し、モデルを作る時点で事前に注意するだけでなく、モデルを作った後に、センシティブな属性が異なるグループごとにAIの判断を比較して、統計的な検定などによって差がないことを確認することが必要になる。また、運用中に追加的に学習してゆくシステムに対しては、運用中にも適切なタイミングでモニタリングを行う必要もある。そして、もしも偏りが検出された場合には、その原因となったデータを同定して、その影響力を減らすようにするなど、バイアス除去（de-bias）の対策を迅速に取る必要がある。バイアスの原因を理解し、修正するために、説明可能AIの技術が利用されることもある。

ここで、どのような属性についてモニタリングを行うか、どの程度のバイアスまでを許容するか、などは、システムの機能や利用の社会的文脈によって決める必要がある。

また、バイアスを除去することは、結果として、予測精度などのシステムの性能を低下させる可能性があるため、その影響についても考慮する必要がある。モデルの学習時に、公平性を担保しつつ、できるだけ精度を落とさないように学習を進める学習アルゴリズムなども提案されている。

公平性を考慮した機械学習モデルの作成、管理、運用を支援するためのツールも少しずつ提供されるようになっている。例えば、IBM社はAI Fairness 360[14]というモデルの作成時にバイアスを緩和するツールや、Open Scale[15]というモデルの運用時にバイアスを監視し、修正するためのツールを提供している。この他にも、Google社のWhat-If Tool[16]、Microsoft社のFairlearn[17]、Amazon社のSageMaker Clarify[18]など、機械学習AIシステムの構築と運用をサポートするための環境が整備されつつある。

ここまでに述べたセンシティブ属性に対するAIの出力の公平性以外にも、例えば、顔認識システムの精度が少数の人種に対して悪化することや、機械翻訳システムの精度が言語によって変わることも、マイノリティーのユーザに対する不公平と考えられる。そうしたバイアスを避けるために、例えば大量の学習データを人工的に生成するような研究[19]も始まっている。

AIシステムがもたらす価値を、幅広い多様なユーザが公平に享受できるように、AIの公平性、バイアスへの配慮と対処は、技術的な側面と、社会的な側面との両面から、継続的に取り組む必要があり、例えば人工知能学会倫理委員会、等の団体が2019年12月に「機械学習と公平性に関する声明」[20]を出している。

4. AIの品質保証

製品やサービスを提供する場合に、品質の評価や管理が重要であることは論をまたない。特に、自動運転、プラントや製造装置の制御、診断や手術の支援、等の実社会での物理的な操作につながるシステムは、事故などによるリスクが大きなものになる可能性があることから、品質管理の重要性が高い。品質保証があることで、ユーザは安心してシステムを利用することができるし、健全な取引契約が可能になる。

AIシステムの中核的構成要素はソフトウエアであり、従来からソフトウエアの品質管理の手法は存在するが、機械学習を用いるAIの場合、従来のソフトウエアとは異なり、その作成過程に、データから学習させることが含まれる。そこでは、プログラムが仕様通りに正しく、効率良く動作するだけでなく、学習の元となるデータの品質、それを使って学習する際のモデルの品質、さらには、システムを運用する際の品質、などがシステムの性能に影響を及ぼすため、従来のソフトウエアの品質評価、管理の方法だけでは十分に対応できない。

また、現在のAIシステムが扱う入力は、画像、音声、テキストなど、大きなバリエーションを持ち、そのバリエーションに対処するために、決定論的ではなく確率統計的な動作を含む機械学習手法が用いられることから、例えば画像認識の精度が100％になることはほとんどなく、人間のように誤りを犯す可能性が常に残る。

さらに、要求される品質についても、システムが行う予測や識別の精度のような「性能」だけではなく、本稿で述べるような、説明可能性、公平性、学習に用いるデータに関する情報セキュリティ、プライバシーの保護、さらには、主に学習時に必要となる電力エネルギー、それに伴って排出されるCO_2量なども、AIシステムに要求される品質の一部となり得る。

こうしたさまざまな課題がAIシステムの品質評価や保証を難しくしており、AIの社会への導入を阻害する可能性があることから、これらの課題に対処して、高い品質のAIシステムを安定的に作成するための、機械学習を用いるAIの品質保証の枠組みと手法についての検討が進められている。例えば、AIプロダクト品質保証コンソーシアム（QA4AI）は、2019年5月に「AIプロダクト品質保証ガイドライン」[21]の初版を発表し、2021年9月には最新版をリリースしている。また、産業技術総合研究所のデジタルアーキテクチャ研究センターを中心とするグループは、2021年6月に「機械学習品質マネジメントガイドライン」[22]の第1版を、2021年7月に第2版を公開している。

「AIプロダクト品質保証ガイドライン」では、品質保証の軸を、Data Integrity（データに関する軸）、Model Robustness（機械学習モデルに関する軸）、System Quality（システム全体に関する軸）、Process Agility（開発・運用プロセスに関する軸）、Customer Expectation（ステークホルダー、顧客、標準などに関する軸）に整理して、それぞれについてチェックリストを提示するとともに、音声インターフェース、産業用システム、自動運転、OCRといった具体的な応用システム種別に即して上記の軸を適用している。さらに、品質保証に利用できる技術のカタログも用意している。

「機械学習品質マネジメントガイドライン」では、AIシステムの品質を、システム全体で最終的な利用者に提供すべき「利用時品質」、システムおよびその構成要素に対して要求される、客観的な視点の品質である「外部品質」、構成要素を構築する際に具体的に測定したり、設計などの開発行為を通じて評価したりできる、その要素が固有に持つ特性としての「内部品質」、に分けて詳しく検討している。具体的には、外部品質の軸として、安全性・リスク回避性、AIパフォーマンス（有効性）、公平性、を取り上げて、さまざまなシステム用途におけるそれぞれの要求レベ

ルを整理するとともに、内部品質管理の特性軸として、品質構造・データセットの設計、データセットの品質、機械学習モデルの品質、ソフトウエア実装の品質、運用時の品質、を抽出し、それぞれの要素特性について、外部品質の要求レベルに対応した内部品質の要求レベルを与えている。

また、ガイドラインに沿った品質指標の測定・検査・改善を支援するツール群と、その作業全体を統括管理できる作業環境として機械学習品質評価テストベッド[23]も併せて構築中である（図表5）。

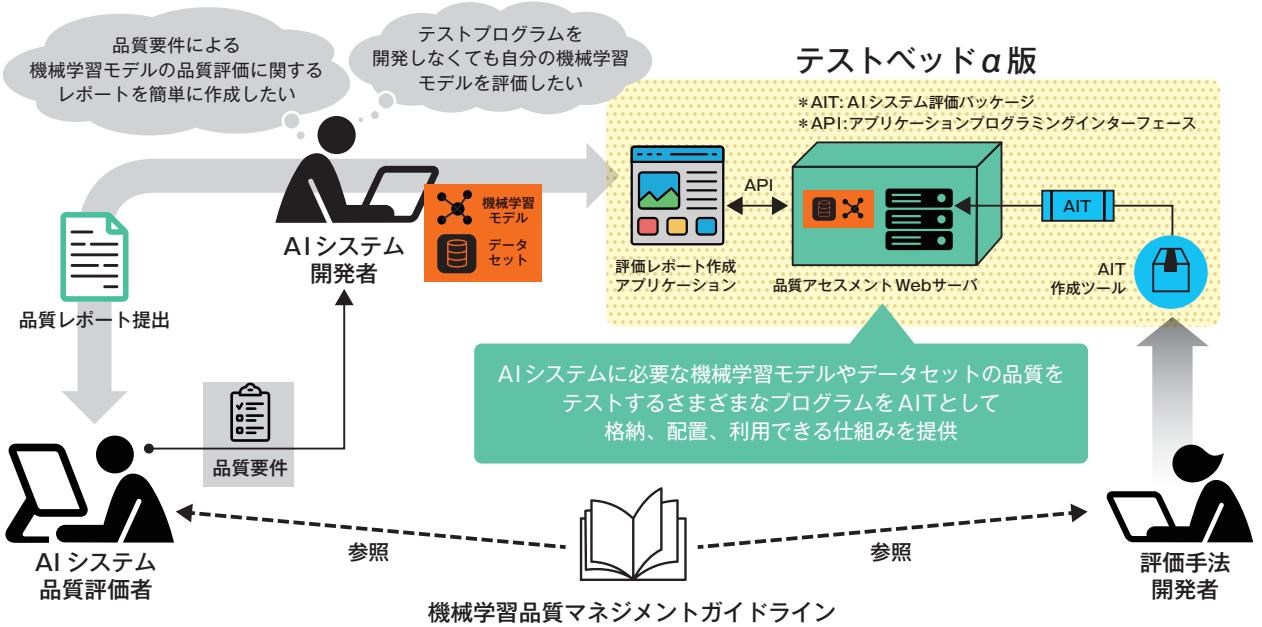
こうしたガイドラインを参考として、個別分野に関するガイドラインを作成する動きもある。例えば、経済産業省、総務省消防庁、厚生労働省の石油コンビナート等災害防止3省連絡会議は、2021年3月に「プラント保安分野 AI 信頼性評価ガイドライン（第2版）」[24]を作成している。また、AI システムの開発やデータの利用に関する契約については、経済産業省の「AI・データの利用に関する契約ガイドライン」[25]などがある。

5. AIのプライバシー保護

現在のAIはデータから知識を学習する機械学習技術に支えられているが、AIを学習するためのデータは、人間の個人情報やプライバシー情報に関わることが多く、それらのプライバシーを保護しつつ機械学習を行うことが技術的課題として検討されている。

プライバシー情報は個人に関するあらゆる情報を指すのに対し、法令における個人情報は法的に保護される範囲を明確にしている。例えば、日本国の個人情報保護法では、個人情報は「当該情報に含まれる氏名、生年月日その他の記述等（中略）特定の個人を識別す

図表5　機械学習品質評価テストベッドα版



出典：産業技術総合研究所プレスリリースを基に作成
（https://www.aist.go.jp/aist_j/press_release/pr2020/pr20201118/pr20201118.html）

ることができるもの」（第2条 第1項1号）を指す。また、「個人識別符号が含まれるもの」（第2条 第1項2号）も個人情報に含まれる。平成29年改正では匿名加工情報が創設され、令和2年改正（令和4年施行）では仮名加工情報が創設され、要件や扱いの異なる個人情報の態様が定義されている。また、令和2年改正では「個人関連情報」も規定され、「生存する個人に関する情報」を含みプライバシー情報全体に近い広い定義となっている。

欧州のGDPR（EU一般データ保護規則）では、オンライン識別子により識別されるデータも個人情報に当たる（第4条第1項）など、日本国法よりも個人情報が広く定義されている。GDPRはEU域内の個人データが含まれる場合に適用され、違反した場合は罰則規定も存在する。また、同様に厳格な法制としてカリフォルニア州のCCPA（カリフォルニア州消費者プライバシー法）がある。

昨今の動向としては、クッキーなどのオンライン識別子に関する商慣行が見直されつつある。日本では、リクナビDMPフォローにおいて、就活生のWeb閲覧行動をクッキーに対応付けてAIにより内定辞退スコアを推定し、企業に情報を提供していた点が問題となった[26]。問題となった点は主に、リクナビ側が提供先企業において個人に対応付けられることを想定しながら提供していた点で、就活生にとってセンシティブな情報が同意なく提供された点であった。このような背景を踏まえ、令和2年改正では、内定辞退スコアのような個人に関連する情報は先述の「個人関連情報」と扱われるようになり、提供先において保有している情報と結び付けることで個人データとなることが想定される情報を提供する場合には、本人の同意が得られていること等の確認が必要（第26条の2第1項）となった。

以上は、法令上の扱いであるが、学術的にはより踏み込んで、悪意の第三者がプライバシーを暴くような攻撃にも対応することも目標とした研究も進められている。一般に知られる技術的な方法としては、1）匿名化、2）差分プライバシー、3）フェデレーテッドラーニング、4）秘密計算がある。

1）匿名化は、入力データにおいて個人を識別しにくいように加工するものである。

匿名化の有名な指針としては、k-匿名性、*l*-多様性がある。k-匿名性[27]は少なくともk人までしか対象を絞り込めないよう加工するという指針で、どの属性を指定した場合も特定の個人は最悪でもk人のうちの一人という一般化された状態を担保する。具体的な操作としては、k-匿名性を満たす上で障害となる属性を排除したり、属性値を丸めるなどの操作を行う。k-匿名性には脆弱（ぜいじゃく）性が知られており、k人の属性が同一であった場合、属性を当てることができてしまい、それが病歴のような要配慮個人情報である場合は問題となる。

l-多様性[28]はこのような課題に対応した指針であり、属性の組み合わせに対して、*l*個の異なる要配慮個人情報を持つように加工するという指針となる。これにより、属性で絞り込んでも要配慮個人情報は*l*個のどれかとなり、要配慮個人情報が特定されにくくなる。

l-多様性にも脆弱性はあり、要配慮個人情報に関するあるグループのデータの分散とデータ全体の分散に違いがある場合、これを手掛かりに確率的な推測が可能になる場合がある。これに対応したt-近似性[29]という指針も提案されている。しかしながら、これらの加工の度合いを高めると、AIが学習する上で情報が失われてしまうという欠点もある。匿名化のようなデータ公開に関する対処法は、「プライバシー保護データ公開（Privacy Preserving Data Publishing）」と整理されることもある[30]。

2）匿名化がデータそのものを加工して公開するのに対し、差分プライバシーは問い合わせの応答において個人を識別できないようにノイズを加えることで対応する。

差分プライバシーの定義は“Differential Privacy”[31]に譲るが、データへの問い合わせに対応し、対象のXさんを除いたデータでも、含むデータでも出力に差異が見えないようにするというのが本質である。これによりXさんの情報が出力から判断できないようにするという指針である。具体的な操作としては、この指針を満たすようAIによる出力過程に対しノイズを加えるという手法が一般的である。Apple社でも、プライバシー対応として差分プライバシーを利用していることが表明されている[32]。

3)フェデレーテッドラーニングは[33]、個人や企業などのローカル環境に分散的に存在するプライバシーに関する生のデータを収集せず、学習を行う方式である。

生のデータに代えて、生のデータから学習したAIモデルのパラメータや学習過程の情報を中央のサーバで収集し、全体AIモデルを更新し、更新後のAIモデルをローカル環境に配布する。生のデータに含まれるプライバシー情報をそのまま送らないため、情報漏えいが抑制できる。データ収集のための通信コストが節約できる可能性もがあるが、AIモデルのパラメータ数や学習の繰り返し回数が多いケースでは、その分のAIモデルの収配信に関する通信コストがかかるため、節約効果は相殺される可能性がある。

また、フェデレーテッドラーニングでは、各ローカル環境でのデータが標準化されていることが暗黙の前提であることが多く、標準化されていない多様なデータに対しても、適用できる手法は今後の課題である。

フェデレーテッドラーニングの実行環境としては、PySyft[34]、TensorFlow Federated[35]、STADLE(TieSet社)[36]など専用ライブラリーも出てきている。また具体的な応用事例として、Google社の予測キーボードにおけるフェデレーテッドラーニングが有名であり、医療分野でのOwkin社[37]など領域特化のフェデレーテッドラーニングのスタートアップ企業も出てきている。

4)最後に秘密計算は、暗号化した状態で計算を行う方式であり、暗号化した状態で扱うため、情報漏えいしづらくなる効果がある。分散環境における秘密計算方式としては、秘密分散方式[38]と準同型暗号方式[39]があり、AIへの適用が進んでいる[40]。

図表6　ユースケースとそれに対応した技術観点の例

ユースケース	概要	説明可能性	公平性	品質保証	プライバシー
推薦システム	Webユーザに商品推薦	推薦理由をユーザに説明	属性でのバイアス排除	新商品などへの対応	オトリアイテムによる漏えい対応
診断支援システム	各地域病院で連携して診断	診断根拠の提示	属性や地域でのバイアス排除	季節・新規状況への対応	要配慮個人情報の保護

以上の2〜4)は、データの処理における対処法であり、「プライバシー保護データマイニング(Privacy Preserving Data Mining)」とまとめられることもある[30]。

6. おわりに

ここまで、AIに対する倫理的なリスクへの技術的な対応として、説明可能性、公平性、品質保証、プライバシー保護に関する技術を解説した。まとめとして、図表6にいくつかのAIのユースケースについて四つの技術観点の対応例を示す。例えば、推薦システムの例では、ユーザの行動データを学習データとして用いるため、そこに含まれるプライバシー情報の保護や、ユーザ属性に関するバイアスの除去が重要になる。また、医療診断支援では、多様な状況への対応を含めた品質保証や診断根拠の説明が重要になる。このように、実際のビジネスユースケースでは、各観点について複合的に対応することが求められる。

本稿で紹介した技術課題の難しさは、突き詰めると、現在のAIが人間とも従来のソフトウエアとも異なる動作原理で動作しているところに由来していると考えられる。現在のAIは、すでに十分複雑であるが、まだまだ未熟なところもあり、記号、離散知識、因果構造などをまだまだうまく扱えてないといわれている。人間の脳と対比すると、システム1と呼ばれる、人間が直観的に素早く処理できる部分は代替できるようになっている一方、システム2と呼ばれる、論理的な思考などのじっくり時間をかけて処理する部分はまだまだ代替できていない[41]。このようなAIの本質的な課題が解決され、より人間に近いAIが実現されていくことで、説明可能AIなど追加的な機構が徐々に必要なくなり、AIが自律的に人間と協調して適用リスクを低減できるようになっていくことが期待できる。その意味で、本稿で紹介した技術群はAIの発展途上における経過的な技術群である可能性もあり、AIそのものの技術的発展も含めて今後の動向を見守る必要があるだろう。

Mori Kurokawa

黒川 茂莉

株式会社 KDDI総合研究所 AI部門
2007年慶應義塾大学院理工学研究科開放環境科学専攻修士課程修了。同年KDDI(株)入社。以降、(株)KDDI総合研究所にてAI、機械学習等の研究に従事。現在、同社統合機械学習グループリーダー。2011年電子情報通信学会学術奨励賞受賞。

Hideki Asoh

麻生 英樹

株式会社 KDDI総合研究所 AI部門
国立研究開発法人 産業技術総合研究所 人工知能研究センター
1983年4月通商産業省工業技術院電子技術総合研究所入所。2015年5月から(国研)産業技術総合研究所人工知能研究センター副研究センター長を務め、現在は同センターの招聘研究員。2019年4月から(株)KDDI総合研究所招聘研究員を兼務。ニューラルネットワークをはじめとする統計的機械学習の基礎理論・アルゴリズムと、学習能力を持つ知的情報処理システムへの応用に関する研究開発に従事。

参考文献

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” In CVPR, 2016.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” In NAACL, 2019.

[3] AIガバナンス（METI/経済産業省），https://www.meti.go.jp/policy/it_policy/ai-governance/index.html

[4] S. Herculano-Houzel, B. Mota, R. Lent, “Cellular scaling rules for rodent brains,” Proc Natl Acad Sci U S A. 2006 Vol. 103, No.32, pp.12138-12143.

[5] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, R. J. Filho, R. Lent, and S. Herculano - Houzel, “Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled - up primate brain,” Journal of Comparative Neurology, Vol. 513, No. 5, pp. 532-541, 2009.

[6] https://keras.io/ja/applications/

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” In ICCV, 2017.

[8] P. W. Koh, and P. Liang, “Understanding black-box predictions via influence functions,” In ICML, 2017.

[9] J. De Fauw et al., “Clinically applicable deep learning for diagnosis and referral in retinal disease,” Nature Medicine, Vol. 24, No. 9, pp. 1342-1350, 2018.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?” Explaining the predictions of any classifier,” In ACM SIGKDD, 2016.

[11] S. M. Lundberg, and S. I. Lee, “A unified approach to interpreting model predictions,” In NeurIPS, 2017.

[12] X. Liu, X. Wang, and S. Matwin, “Improving the interpretability of deep neural networks with knowledge distillation,” In ICDM Workshops, 2018.

[13] 富士秀 , 森田一 , 後藤啓介 , 丸橋弘治 , 穴井宏和 , 井形伸之 , “Deep Tensor とナレッジグラフを融合した説明可能な AI（特集 デジタル革新を支える人工知能 Zinrai）―（Zinrai を支える最先端テクノロジー）” Fujitsu, Vol. 69, No. 4, pp. 90-96, 2018.

[14] AI Fairness 360，https://aif360.mybluemix.net/

[15] Watson Open Scale，https://www.ibm.com/jp-ja/cloud/watson-openscale/model-risk-management

[16] What-If Tool，https://pair-code.github.io/what-if-tool/

[17] Fairlearn，https://fairlearn.org/

[18] SageMaker Clarity，https://aws.amazon.com/jp/sagemaker/clarify/

[19] MIT Technological Review，https://www.technologyreview.jp/s/234371/fractals-can-help-ai-learn-to-see-more-clearly-or-at-least-fairly/

[20] 機械学習と公平性に関する声明，http://ai-elsi.org/archives/888

[21] AIプロダクト品質保証ガイドライン（2021.09版），http://www.qa4ai.jp/QA4AI.Guideline.202109.pdf

[22] 機械学習品質マネジメントガイドライン（第2版）
https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev2.html

[23] 機械学習システムの品質評価テストベッドα版（機能限定），https://github.com/aistairc/qunomon

[24] プラント保安分野 AI信頼性評価ガイドライン
https://www.meti.go.jp/press/2020/03/20210330002/20210330002-2.pdf

[25] AI・データの利用に関する契約ガイドライン
https://www.meti.go.jp/policy/mono_info_service/connected_industries/sharing_and_utilization.html

参考文献

［26］ 工藤郁子 , 荒井ひろみ , 江間有沙 , “採用におけるプロファイリング・サービスの倫理的課題「リクナビ DMP フォロー」事件を手がかりに ,” 人工知能学会全国大会論文集 第 34 回 , 2020.

［27］ L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 05, pp. 557-570, 2002.

［28］ A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” ACM TKDD, Vol. 1, No. 1, pp. 3-es, 2007.

［29］ N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” In ICDE, 2007.

［30］ 高橋克巳 , “プライバシー保護データマイニング”, システム / 制御 / 情報 , Vol. 63, No. 2, pp. 43-50, 2019.

［31］ C. Dwork, “Differential Privacy,” In ICALP, 2006.

［32］ Apple Inc., https://www.apple.com/jp/privacy/features/

［33］ M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, “Federated learning: A survey on enabling technologies, protocols, and applications,” IEEE Access, Vol. 8, pp. 140699-140725, 2020

［34］ PySyft, https://openmined.github.io/PySyft/

［35］ TensorFlow Federated, https://www.tensorflow.org/federated

［36］ TieSet STADLE, https://www.tie-set.com/services

［37］ Owkin, https://owkin.com/

［38］ A. Shamir, “How to share a secret,” Communications of the ACM, Vol. 22, No. 11, pp. 612-613, 1979.

［39］ C. Gentry, “Fully homomorphic encryption using ideal lattices,” In Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009.

［40］ 三品気吹 , 濱田浩気 , 五十嵐大 , 菊池亮 , “秘密実数演算を用いた高速かつ高精度なロジスティック回帰とデータ標準化 ,” CSS, 2020.

［41］ Yoshua Bengio, “Deep Learning for System 2 Processing presentation,” the AAAI-20 Turing Award winners 2018 special event, February 9, 2020, http://www.iro.umontreal.ca/~bengioy/AAAI-9feb2020.pdf

取材リポート

5年後の未来を探せ

近藤 能子　長崎大学 大学院 准教授に聞く

気候変動を左右する海洋の微量金属元素を追う

取材・文：江口絵理　撮影：八木拓也　図版提供：近藤能子（Figure2、3を除く）

2021年のノーベル物理学賞は、地球温暖化予測のための気候モデルを開発した真鍋淑郎さんらに授与された。地球の気温は大気だけでなく海洋にも大きな影響を受けるため、モデルを使って正確に地球の将来を予測するには、実際の海洋のデータが欠かせない。長崎大学大学院水産・環境科学総合研究科の近藤能子准教授は、「化学海洋学」を軸足に、二酸化炭素吸収源でもある海で重大な役割を果たすミクロな物質を追いかけ、基礎的なデータと貴重な知見を積み上げている。

世界の海を舞台に、ごく微量の物質を探す

ある年は日本の港から出発して北極海へ。ある年にはインド洋へ。ニュージーランドを出港し、赤道直下の海域を経て日本へ戻ってくる航海もあれば、南米チリからハワイまで南太平洋を北上する航海にも参加した。壮大なスケールの研究航海を重ねる近藤能子さんが追っているのは、フィールドのスケールの大きさとは対照的に、目に見えないほど小さな物質。海水に溶け込んでいる金属だ。

「鉄をはじめ、マンガンやニッケル、亜鉛、コバルト、カドミウムなどの微量元素を対象として研究しています。『世界の海をフィールドに』といえば華々しく聞こえますが、海水中にごくごく微量しか存在していない元素なので、ただ捕集するだけにも想像以上に骨が折れる、地味な仕事です」と近藤さんは笑う。

海水中に含まれる量が少ないならなおさら、サンプルにほんのわずかでも鉄粉やさびなどが混じれば結果がまるで変わってしまう。ところが近藤さんが乗る船は、ほとんど鉄の塊だ。

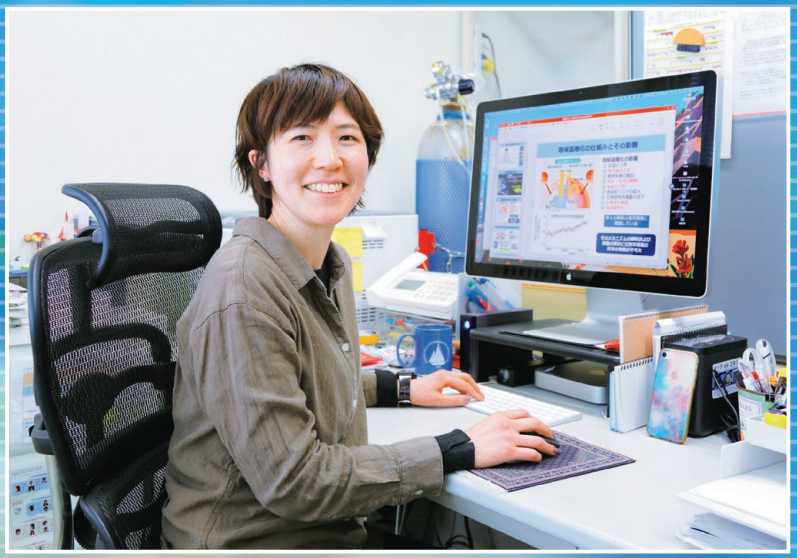
「サンプルを採取するボトルを海中から引き上げ、船上で海水試料を取り出すときは、船体から鉄を含むものが舞い落ちてきて混入しないよう、甲板に密閉したブースを作り、採取者はその中でプラスチックの手袋をはめて作業します」

溶け込んでいる金属は極めて微量なので、時には自ら工作した特殊な器材を使って海水を濃縮し、元素別に濃度（海水に溶け込んでいる量）を分析。それを、採取場所を変えて採水するたびに行うという。

それほどまでに採取や分析に手間がかかり、ごく微量しか海水に含まれていない金属を、近藤さんはなぜ研究対象に選んだのか。

「実は、鉄などの微量金属は、海洋の生物生産の基盤となる植物プランクトンに欠かせない存在なのです」

植物プランクトンは、陸上の植物と同様、光合成によって自身の体やエネルギーを作り出す。植物プランクトンの成長と増殖には、二酸化炭素と水、日光だけではなく、窒素やリン、珪藻類ではそれに加えてケイ素が必要だ。一方、鉄をはじめとする金属は、必要な量こそはるかに少ないものの、「窒素やリンなどの代謝を助ける」という重要な役割を果たしている。



Yoshiko Kondo

近藤 能子

2007年、東京大学大学院 農学生命科学研究科 水圏生物科学専攻 博士課程修了。長崎大学 水産・環境科学総合研究科 助教、同 海洋未来イノベーション機構 助教などを経て2018年より現職。2017年、「長崎大学未来に羽ばたく女性研究者賞」【優秀女性奨励賞】、海洋化学奨励賞受賞、2019年、資生堂女性研究者サイエンスグラント受賞。

「例えば鉄は、光合成や葉緑素の生合成、窒素の代謝などに植物プランクトンが必要とする元素です」

そのため、植物プランクトンの「主要栄養塩」と呼ばれる窒素、リン、ケイ素が周囲にふんだんに存在していても、鉄が足りなければ植物プランクトンは成長・増殖できない。そして、海洋における食物連鎖のピラミッドを最も下で支える植物プランクトンの多寡は、ピラミッド上層の海洋生物にダイレクトに影響を与えている。

「微量といえども鉄やマンガンなどの微量金属は、海洋の生態系と、海がもたらす水産資源を利用する私たち人間に巨大な影響を与えているんです」

植物プランクトンは地球温暖化も左右する

さらに、植物プランクトンは地球の気候変動にも大きな役割を果たしている。「近年の地球温暖化は人間活動に由来する二酸化炭素排出量が増大し、森林など、地球上の二酸化炭素吸収源のキャパシティーを超過していることが原因ですが」と前置きして、近藤さんは続けた。

「二酸化炭素の吸収源と聞くと、熱帯雨林を思い浮かべる方が多いでしょう。確かに陸上の森林が吸収する量は多いのですが、ある試算によると、植物プランクトンなどによって海洋が吸収する割合は、地球全体の吸収量の50%にも上るといわれています」

すなわち、植物プランクトンの繁栄を支える栄養素について知ることは、人類が抱える喫緊の問題と密接につながっている。そのため、主要栄養塩については、これまでも世界中の海洋学者が精力的に研究を進めてきた。

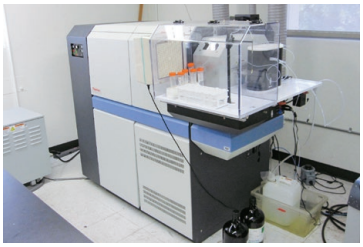
「海洋調査の蓄積や人工衛星による観測技術の発達によって、まだまだ粗い部分はあるにせよ、世界中の海の栄養塩の分布を全球マップで描くことができるまでになっています」と、近藤さんは硝酸塩の分布図を指し示した。

「ただ、このマップで不思議なのは、海洋のほとんどの場所で窒素の供給源である硝酸塩が使い尽くされている一方で、局所的に硝酸塩が余っている海域があることです。植物プランクトンは普通、栄養塩を使い尽くすまで増殖を続けるので、南極周辺や北海道沖のように硝酸塩が余っている海域には、植物プランクトンが増えられない足かせが別に存在すると考えられます」

しかもそうした海域は全海洋の30%もの割合を占めている。海洋学では、その“犯人”候補の一つが鉄だといわれてきた。それが、近藤さんを含め、海洋環境の研究者が微量金属に着目する理由である。

しかし海中の鉄の量を確かめようにも、その量は海水1kg当たり6×10^{-10}～10^{-8}g程度と、主要栄養塩と比べて3桁も少ない。海水から捕集する際の異物混入を防いで、微量金属の正確な量を測る技術がそれなりに

Figure1　調査の様子



調査をする海域で採集した海水からサンプルを採取し分析する。対象が主に微量金属であるため、サンプル採取に際しては、船体などの金属を混入させないよう、慎重な作業が求められる。サンプルは濃縮した後に調査対象の金属の質量を分析する

確立してきたのは1980年代以降で、人工衛星から観測するすべもなく、近藤さんを含む世界の研究者たちが広大な海洋で一つ一つサンプリングを積み重ねている状況だ。

しかも、海の中で、鉄は局在している。海水に溶け込んでいるさまざまな物質と同様、鉄も陸地から多く供給されるため沿岸域では濃度が高く、また海底の割れ目から地殻由来の鉱物が熱水と共に噴出している熱水鉱床と呼ばれるエリアでも濃度が高いが、海水に溶けにくいため表層域の鉄の量は多くない。海底にも多く堆積していて、堆積物が舞い上がるところでも濃度は高くなる。しかしこうした特徴的な例を除き、鉄をはじめとする微量金属の分布と移動のメカニズムには未知の部分の方がはるかに多い。

金属ごとに、海中での振る舞いが異なることを発見

微量金属の移動と分布のメカニズムを実際に確かめられたのが、2012年と2013年の2回にわたり、夏の北極海で行われた調査だ。北極海は毎年、春夏になると海氷が融解し植物プランクトンが大発生して生物生産が極めて盛んになる。近藤さんらが注目したのはユーラシア大陸とアメリカ大陸の切れ間に当たるベーリング海峡。より正確には、海峡の北に広がる「チャクチ海」と呼ばれる海域だ。

「ベーリング海峡では太平洋側から北極海に向かって海水が一方通行で流れていくので、北極海の入り口に当たるチャクチ海は海水中の物質移動を観測するのに非常に適しているんです。しかも、チャクチ海は陸棚が広がっていて水深が50mほどと非常に浅く、海水は海底堆積物から微量元素の供給を受けていることが予想されます」

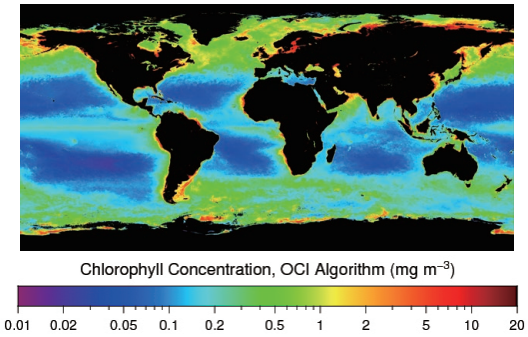
海峡近くの浅い海域からチャクチ海の陸棚域を過ぎると、海底はすとんと落ち込んで、深さのあるカナダ海盆へとつながっていく。この航海でチャクチ海からカナダ海盆に向かって5点の観測地点を設定し、それぞれ5mから500mの深さで海水のサンプリングを行った。すると、浅い海域ではリン酸などの主要な栄養も、鉄やマンガンなどの金属も多く含まれ、浅い海底の堆積物から巻き上げられて海水中に多くとどまっていることが明らかになった。

その一方で、リン酸は比較的遠くまで流されていくのか、陸棚域から深い海域まで変わらず高い濃度を保っていたが、鉄は陸棚域より沖合の深い海域では急減していた。

「鉄は海水に溶けにくいので、すぐに海底に落ちてしまうと以前からいわれていましたが、この観測結果はその通説を科学的に補強するものになりました」

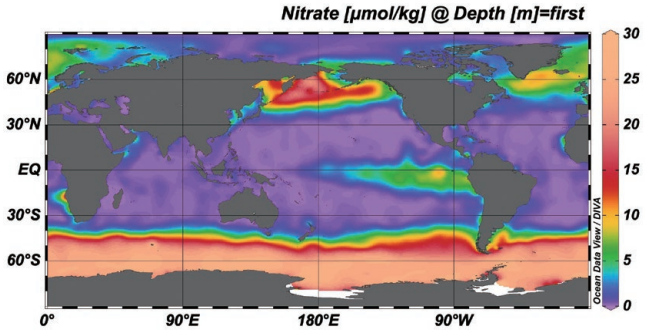
さらに、金属なら何でも鉄のようにすぐに海底に沈降していくわけではなく、同じく植物プランクトンが主要栄養素を代謝するのに必要とする微量金属元素でも、亜鉛などは海水中にとどまり、鉄よりもはるかに遠くまで届いていた。こうした元素ごとの海中での振る舞いの違いを実地に観測できたことも、大きな成果の一つだ。

Figure2　海洋表層のクロロフィルａ濃度（年平均）



クロロフィルａ濃度は植物プランクトン現存量の指標。大陸周辺には多く存在しているが、沖合にはほとんど存在しないことが見て取れる（2021 MODIS-Aqua Chlorophyll concentration, https://oceancolor.gsfc.nasa.gov/l3）

Figure3　海洋表層の硝酸塩濃度（年平均）



硝酸塩濃度は植物プランクトンの成育に欠かせない窒素濃度の指標。多くの海域で硝酸塩はほとんど残っていない一方で、豊富な海域もあることが分かる（World Ocean Atlas（2018）, https://www.ncei.noaa.gov/products/world-ocean-atlas）

海水中の鉄が取る多彩な「化学的形態」

あらためて考えてみれば、地球において、鉄はごくありふれた物質だ。地殻には鉄が大量に含まれ、陸上にも鉄は多く存在する。海水中にごくわずかしか存在しないのは、鉄が海水に溶けにくいからだ。とりわけ、植物プランクトンが利用しやすい無機鉄として溶け込んでいる量は極めて少ない。

では植物プランクトンは、絶対量の少ない無機鉄を利用できる範囲内で増殖しているかというと、そうでもない。別の“形”の鉄も利用しているのだ。

「鉄は、海水中のさまざまな有機物、例えば生物の死骸や糞（ふん）、食べかすなどと結び付いた化合物として存在することもできるんです。これを『有機錯体鉄』といいます。有機錯体鉄は無機鉄より海水に溶けやすく、従って海水中の量は無機鉄より多い。また無機鉄が凝集した『コロイド鉄』という状態もあります」

急に化学の話になったように見えるが、実はこれが近藤さんの専門。近藤さんは「化学海洋学」の研究者なのである。

同じ「鉄」でも、無機の鉄として存在しているのか、有機物と結び付いている錯体なのか、あるいはコロイド鉄なのか、つまりどのような化学的形態であるかによって、海水への溶解度や生物にとっての利用しやすさは異なるという。

「有機錯体鉄やコロイド鉄は無機鉄よりも海水中に多く溶け込めるので、植物プランクトンにとって重要な鉄のソースです。けれども、最終的に取り込みたいのは鉄だけなので、有機物との結合を切り離すのにはエネルギーが必要になります。使いやすい無機鉄が足りないから、植物プランクトンは使いにくい形態の鉄も利用せざるを得ない、というのが実情だと思います」

海水中の鉄の総量だけを見ていたのでは分からないことがある。近藤さんたちの研究はそれを明らかにしつつある。

「ただ、鉄と結び付くことのできる有機物の種類は数多く、さらに、分解などを経て、私たちがこれまで見たことのない未同定の化合物として現れることもあります。何に由来するどんな有機物なのか、正体を突きとめるのもなかなか大変です」

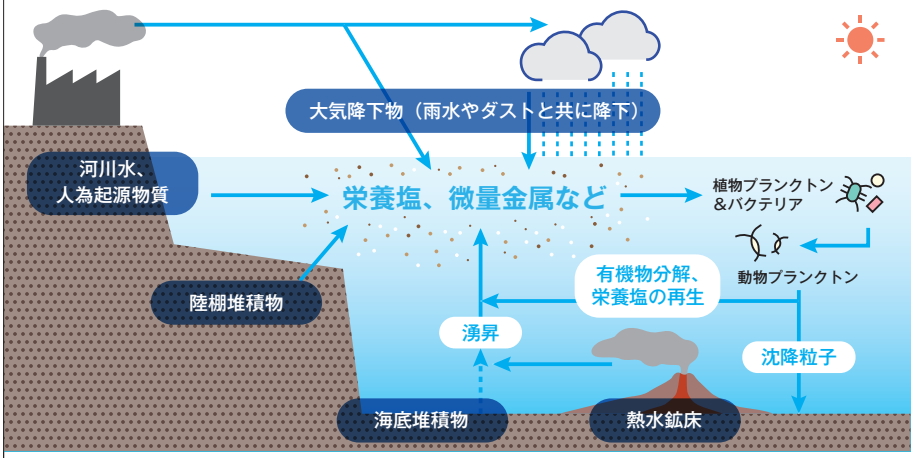
2021年、近藤さんたちのチームは、亜寒帯太平洋の西と東では有機錯体鉄の量に違いがあることを示す論文を発表した。

「表層はどこも鉄の濃度が低いのですが、深層では、西の方が濃度が高く、東側は低いのです。それはいったいどうしてなのか、海洋学における長年の疑問でした」

近藤さんたちは、鉄と結び付き得る有機化合物の量が東西で違うのではないかという仮説を基に調査を行ったところ、果たして西側では、有機錯体鉄やコロイド鉄の量が東側より多いことが分かった。鉄の化学的形態の多彩さが海域ごとの鉄の量の差をもたらしている例を示したこの成果は、まさに「海の化学屋」の仕事といえる。

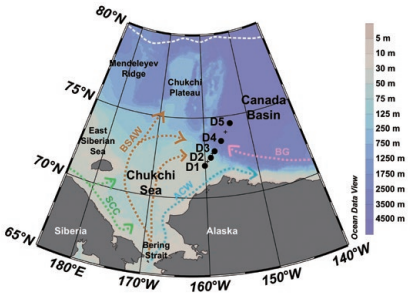
有機錯体鉄の捕集や分析は、鉄だけを扱うよりも難

Figure4　海洋における微量物質循環の模式図



河川から流れ込んだり大気から降下した微量物質は、海底から海面に向かう湧昇流、潮流などによって移動し、さらに海洋中の化学反応や生物活動によって姿を変えながら循環していく

Figure5　北極海での調査を行った海域



ベーリング海峡の北側にあるチャクチ海での調査海域。右の指標は水深で、D1～D5がサンプルを採取した場所（Kondo et al., 2016）

しく、データの蓄積がまだ乏しい。このような化学的アプローチは、海洋中の鉄の局在の背景を探る強力な方法になるかもしれない。

でも、何か一つ分かると、分からないことがさらに増える、と近藤さんは苦笑いする。

「有機錯体鉄を作れる有機物が亜寒帯太平洋の西側に多いことは分かった。すると、なぜ西側に多いのか、という疑問が生まれますよね。そうした有機物を含む水塊はどこから、どのようにここへ移動してくるのだろう？　一つの答えが次の謎を呼ぶという具合で、掘っても掘ってもキリがないんです（笑）」

なぜ、東側には鉄と結び付く有機物が供給されないのか。そもそも、沿岸から流れ込む鉄はどのように海中を移動していくのか。また、陸地や海底などから海中に入ったものの溶け切れなかった無機鉄は、海底に沈降した後にどうなるのか。確かに、多くの実測値が入手できても、データをスナップショットとして見ているだけでは全体像は描けない。

親潮や黒潮に代表される表層の海流のほかに、海洋深層にはまた別の海流がある。地球の海を一巡するのに1000年以上かかるほどのゆっくりした流れに乗って、海底堆積物や熱水鉱床由来の鉄も少しずつ動いているかもしれない。

あるいは極域では冬季に表層が冷えることで表層の海水の密度が上がり、深い方へと沈んでいく（その代わりに深層にあった海水が表層に上がってくる）垂直方向の海水の循環もある。そうした水塊の動きはどれほど海底の鉄を巻き上げるものなのか。

水温や密度、海流や地形など、さまざまな要素が複雑に組み合わさって海水が動き、そこに溶け込んだ物質が地球を循環していく。その過程で鉄が極端に少ない海域ができることもあれば、鉄が豊富に供給されるホットスポットもできる。そうした局在の背景にあるメカニズムを知ることが、近藤さんたち海洋学者が取り組む大きなテーマだ。

「こうなってくると、私が専門とする化学だけではとうてい太刀打ちできません。物理学、地学、生物学の研究者とも手を携えて研究して、初めて全体像に迫れるようになります」

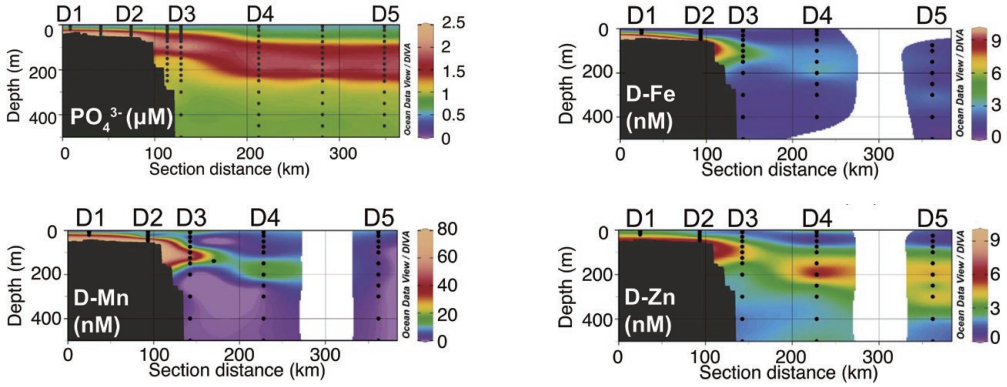
近藤さんは「少なくとも私が見る範囲では」と前置きをしてこう続ける。

「だから、海洋環境研究の世界は、一つのホットトピックにみんなが群がって成果を競い合うのではなく、それぞれが得意な分野で、足りない知見を埋めていこうというマインドが強いんです。競争よりも連携が、大きな謎を解くための近道だという考え方です」

とりわけ多くの分野の知見を集めて大きな成果を出せるのが、海洋のモデルを作る分野の研究だ。この分野では昨年、地球の気候変動モデルの基礎ともいえる大気海洋結合モデルを作った真鍋淑郎博士らのノーベル物理学賞受賞があり、二酸化炭素排出量過多による地球温暖化という深刻な問題を抱える社会からの期待も熱い。

今後、地球環境がどうなっていくのか、科学者にはより正確に予測していくことが求められる。海洋環境をシミュレートするモデルはそのための重要なツール

Figure6　北極海調査の分析結果



調査で得られた窒素および微量元素の濃度。縦軸は水深で横軸D1～D5が採取地点。リン酸（PO_4^{3-}）は調査海域全体で濃度を保っているのに対して、鉄（D-Fe）やマンガン（D-Mn）は、あまり遠くまで流されておらず、深い海にもあまり到達していない。一方で、同じ微量金属でも亜鉛（D-Zn）はより遠く、深くまで濃度を保っている（Kondo et al., 2016を一部改変）

だが、モデルを精緻化するには、正確な実測データの蓄積が欠かせない。近藤さんの研究はそれを化学の面から支えている。

長崎の海の環境課題にも専門を生かして貢献したい

近藤さんは近年、自身が拠点とする長崎大学の近くの海でも調査を行っている。

「長崎には、干満差の大きさで知られる有明海や、貧酸素化が顕著な大村湾など、非常に特徴的な海があります。このような近くの海で頻度高く観測すると、これまで私が遠洋に出かけていって、そのとき限りの調査で見てきたものとはずいぶん異なる物質輸送の様が観察できるんです。中でも物質の分布や移動が潮汐の影響をどう受けるかを見られるのが興味深いです」

有明海や大村湾のような沿岸の閉鎖性海域では、赤潮の発生が深刻な問題となっている。近藤さんは、赤潮を起こす植物プランクトンである渦鞭毛藻類や珪藻類の中には、代謝に必要なはずのビタミンB_{12}が海中に存在しなくても増殖できるものがいることに注目した。

「赤潮の頻発という問題を解決するには、まずはそこで植物プランクトンがどのように生きているのかを知ることが大事ですよね。不思議なことに、彼らは実験室だとビタミンB_{12}を与えないと増えないのに、自然環境だとなくても増えていきます。これは、自然環境では植物プランクトンにビタミンB_{12}を供与する生物がいるなどの可能性を示唆しています。そこでまずは、海中のビタミンB_{12}の分布や輸送について調べています」

実は、ビタミンB_{12}を海中から集めてくるのも、操作は複雑だ。鉄ならば海水50ミリリットル程度の試料で分析できるが、ビタミンB_{12}ではその40倍の2リットルが必要。当然、分析にも時間がかかる。

しかし「大変なんです」と言いながらも、近藤さんは楽しそうだ。

「自分の研究で地域に貢献することができるなら、それはとてもうれしいことですし、フィールドでの調査・分析には、世界で初めて自分がこの事実を知ったんだ！　という喜びがありますから」

世界の海洋に微量の鉄を探し、地元の海にさらに微量のビタミンを探す。海洋の物質循環を通じて地球というシステムのメカニズムに迫り、その将来を予測するため、さまざまな学問分野と連携して知見を積み上げていく海洋学という学問の懐の深さに、近藤さんは魅せられている。

「今はまだデータが十分とはいえませんが、いずれは、海洋中の鉄をはじめとした多くの微量元素の全球マップも描かれるようになるでしょう」

観察しやすい陸上に比べ、人類はまだ海の姿をほとんど知らない。パズルに小さなピースを一つはめるだけで、見える世界が変わることもあるだろう。近藤さんがほかの分野の研究者と手に手を取って切り出してきた次なるピースが、海洋環境研究や気候変動予測に新たな観点をもたらすかもしれない。

「Nextcom」論文公募のお知らせ

本誌では、情報通信に関する社会科学分野の研究活動の活性化を図るため、
新鮮な視点を持つ研究者の方々から論文を公募します。

【公募要領】

申請対象者：大学院生を含む研究者
＊常勤の公務員（研究休職などを含む）の方は応募できません。

論文要件：情報通信に関する社会科学分野の未発表論文（日本語に限ります）
＊情報通信以外の公益事業に関する論文も含みます。
＊技術的内容をテーマとするものは対象外です。
およそ1万字（刷り上がり10頁以内）

選考基準：論文内容の情報通信分野への貢献度を基準に、Nextcom監修委員会が選考します。
（査読付き論文とは位置付けません）

公募論文数：毎年若干数

公募期間：2022年4月1日～9月10日
＊応募された論文が一定数に達した場合、受け付けを停止することがあります。

選考結果：2022年12月ごろ、申請者に通知します。

著作権等：著作権は執筆者に属しますが、「著作物の利用許諾に関する契約」を締結していただきます。

掲載時期：2023年3月、もしくは2023年6月発行号を予定しています。

執筆料：掲載論文の執筆者には、5万円を支払います。

応募：応募方法ならびに詳細は、以下「Nextcom」ホームページをご覧ください。

その他：1. 掲載論文の執筆者は、公益財団法人KDDI財団が実施する著書出版助成に応募することができます。
2. 要件を満たせば、Nextcom論文賞の選考対象となります。
3. ご応募いただいた原稿はお返しいたしません。

「Nextcom」ホームページ
https://rp.kddi-research.jp/nextcom/support/

問い合わせ先：〒105-0001　東京都港区虎ノ門2-10-4 オークラプレステージタワー
株式会社 KDDI総合研究所 Nextcom編集部
E-mail:nextcom@kddi.com

2022年度 著書出版・海外学会等参加助成に関するお知らせ

本誌では、2022年度も公益財団法人KDDI財団が実施する著書出版・海外学会等参加助成に、候補者の推薦を予定しています。

【著書出版助成】

助成内容： 情報通信に関する社会科学分野への研究に関する著書

助成対象者： 過去5年間にNextcom誌へ論文を執筆された方

助成金額： 3件、各200万円

受付期間： 2022年4月1日～9月10日（書類必着）

【海外学会等参加助成】

助成内容： 海外で開催される学会や国際会議への参加に関わる費用への助成

助成対象者： 情報通信に関する社会科学分野の研究者（大学院生を含む）*

助成金額： 北米東部 欧州 最大40万円　北米西部 最大35万円　ハワイ 最大30万円
その他地域 別途相談（総額100万円）**

受付期間： 随時受け付け

*常勤の公務員（研究休職などを含む）の方は応募できません。
Nextcom誌に2頁程度のレポートを執筆いただきます。

**助成金額が上限に達し次第、受け付けを停止することがあります。

推薦・応募： いずれの助成も、Nextcom監修委員会において審査・選考し、公益財団法人KDDI財団へ推薦の上、決定されます。応募方法ならびに詳細は、以下「Nextcom」ホームページをご覧ください。

「Nextcom」ホームページ
https://rp.kddi-research.jp/nextcom/support/

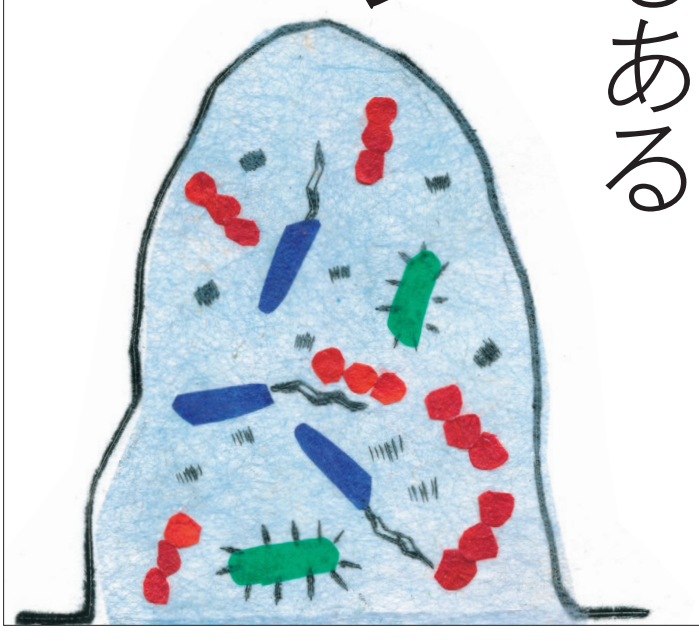
問い合わせ先：〒105-0001　東京都港区虎ノ門2-10-4 オークラプレステージタワー
株式会社 KDDI総合研究所 Nextcom編集部
E-mail:nextcom@kddi.com

彼らの流儀はどうなっている？

執筆：野村 暢彦　絵：大坪 紀久子

微生物は、増えるだけの単純な生物ではなかった。
異種間ですら、多様なコミュニケーション手段を持ち、集団を統制していた。

デジタルシステムもある細菌の多様なコミュニケーション



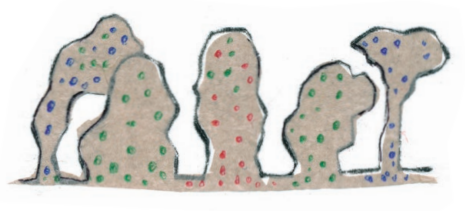
細菌が主（あるじ）で体は入れ物？

目に見えない微生物は「それって生き物？」と思っている方が多いのではないでしょうか。微生物は単細胞の生き物です。大きく分けて、動物・植物などの真核細胞生物に入る真核微生物と細菌、古細菌の三つに分類されます。ここでは、細菌のコミュニケーションについて紹介します。

皆さんがすぐにイメージできるのが、腸内細菌でしょうか。人間の体は約38兆個の細胞から成り立っています。しかし、腸内には40兆から100兆個ものさまざまな種類の細菌が存在しています。人間の体を構成する細胞数より、腸内細菌の細胞数の方がはるかに多いのです。もし、科学の進んだ宇宙人がわれわれを解析すると、腸内細菌が主（あるじ）で、人間をその入れ物と見なすかもしれません。腸内細菌は、免疫機能のみならず、糖尿病や癌、

Nobuhiko Nomura　筑波大学 生命環境系 教授／微生物サステイナビリティ研究センター センター長

1966年生まれ。1995年広島大学大学院工学研究科修了。博士（工学）。
筑波大学生命研究系教授などを経て、2018年より現研究センター兼任。2022年より現職。
JST-ERATO野村集団微生物制御プロジェクト研究総括（2015–2022年）。
専門は微生物生態学、バイオテクノロジー。
微生物細胞の相互作用の多様性・集団性の関係理解をテーマに研究している。



さらに自閉症など脳の機能までと幅広く健康に深く関与していることが明らかになりつつあります。そうした背景から、人間は細菌との複合生命体で、超生命体と呼ばれています。健康においても腸内細菌が重要となり、微生物の制御は21世紀の必須な課題といえます。

細菌は栄養さえあれば増殖するだけの生き物と思われていました。ところが、群れて集団となって生活し、細胞同士でコミュニケーションしながら増殖や機能（有用物質や毒素など）を制御していると分かってきました。

細菌は、自身の細胞から特定の化学物質を外に発信し、それが言葉として用いられます。それを聞いた（取り込んだ）細胞は、言葉に呼応（細胞内の遺伝子発現が制御され、細胞の振る舞いが変化）します。その言葉は化学物質なので、その量（濃度）は細胞から離れるほど減り、遠い細胞は呼応できません。これは近くの細胞同士、つまり言葉が届く範囲のコミュニケーションシステムで、声が届く範囲の細胞が同調するシステムです。

例えば、病原細菌は１匹（1細胞）で、いきなり毒素を出しても宿主の免疫にやられてしまいます。そこで、病原細菌は仲間細胞が増えるのを待って、それを言葉で認識して一斉に同調して毒素を産生し、一気に宿主細胞を攻撃するのです。

デジタルなコミュニケーション発見

さらに私たちは、細胞外膜小胞（MV）と呼ばれる「小包」を介した細胞の「デジタルなコミュニケーションシステム（情報伝達機構）」を発見しました。1個のMVには相互作用に十分な量の言語（化学物質）が濃縮されており、細胞間相互作用に関わる化学物質（言語）の単なる濃度勾配によらない、離散的にかつ細胞特異的に細胞間コミュニケーションが行われるのです。

MVは宛名のついた手紙のように、「届けたい細胞」あるいは「受け取る細胞」を限定できるのです。化学物質（言語）のみを細胞外に放つコミュニケーションは、受け取る細胞が近くにいることが必要条件で、細胞は限定されません。しかし、MVを介したコミュニケーションは細胞が遠く離れていてもコミュニケーションでき、さらに受け取る細胞は選択制で、限定できるのです。興味深いことに実際に海洋や土壌、さらに腸内はMVだらけなのです。

細菌は、さまざまなところで情報をMVによりクラウド化し、情報を「離れた細胞へ」「渡したい細胞へ」「時を越えて細胞へ」渡すことができ、情報を時空間的に制御しています。細菌はこのようなコミュニケーションシステムを地球規模で行っているのです。細菌のコミュニケーションシステムも結構すごいと思いませんか？

よりよくする方法がある。それを見つけよ。
……トーマス・エジソン

人生は勝負だ

髙橋秀実

発明家、トーマス・エジソンの研究室にはこんな標語が掲げられていたという。

「よりよくする方法がある。それを見つけよ」

発明の心得ということか。当たり前といえば当たり前の言葉なのだが、私は瞠目(どうもく)した。これは小学校1年生の「道徳」の教科書に出てくるフレーズにそっくりなのだ。「よりよく　いきるために　たいせつな　ことについて、みんなで　かんがえるよ」*。文部科学省の学習指導要領でも「道徳」は「よりよく生きるため」の教科と規定され、生徒たちはその方法を探るのである。「道徳」とは人生における発明ということか。もしかすると「道徳」教育の原型はエジソンだったのかもしれない。

確かに、偉人伝といえばエジソンというくらいに彼は道徳的である。貧しい家に生まれ、小柄で耳が不自由だったエジソン。級友からいじめられてもへこたれず、都会に出て電信技師として身を立てる。「天才とは1％の閃きと99％の努力」という名言を残すくらいで、彼は「勤勉」「努力」「探究」の人。蓄音機や電灯、電力の配電システムや録音、映画などを次々と発明し、その偉業は「20世紀を発明した」とさえ讃(たた)えられている。エジソンの言う「よりよくする方法」は人生を「よりよく生きる」処世術にも通じていたのか。

真意を探るべく、あらためて原語を調べてみると、「よりよくする方法」の「よりよく」とは英語の「better」だった。

一般的に「better」は「よりよく」「よりよい」などと訳されがちなのだが、これは誤訳ではないだろうか。というのは「better」は比較級の形をしているが、実はもともと原級がない。「better」という語が先にあり、後からこれの原級としてgoodが代用（これを補充法という）されたにすぎないのである。つまりbetterは「より」+「よい」と2語で訳してはいけない。1語でbetterなので、

article：**Hidemine Takahashi**
ノンフィクション作家。1961年横浜市生まれ。東京外国語大学モンゴル語学科卒業。
『ご先祖様はどちら様』で第10回小林秀雄賞受賞、『「弱くても勝てます」開成高校野球部のセオリー』で第23回ミズノスポーツライター賞優秀賞。他の著書に『からくり民主主義』『定年入門』『不明解日本語辞典』『損したくないニッポン人』『悩む人』『パワースポットはここですね』『一生勝負』など。最新刊は『道徳教室　いい人じゃなきゃダメですか』（ポプラ社）。

1語で「すぐれる」などと訳すべきなのだ。漢字なら「優れる」、いや「勝れる」がピッタリではないだろうか。

実際、エジソンは他の発明者との競り合いの末に約1,100件もの特許を勝ち取ったわけだし、そもそも新しい技術というのはそれまでの技術に勝っている。例えばレコードとCDを比べると、良し悪しではそれぞれに一長一短があるが、市場ではCDが勝っている。新しい技術とは「よりよい」ではなく、勝れた技術なのだ。発明とは勝負。「道徳」の「よりよく生きる」も「人生に勝つ」と言ったほうが、気合いも入るのではないだろうか。

*『しょうがく　どうとく　いきる　ちから　1』（日本文教出版　平成30年）

背景
トーマス・エジソン（1847～1931年）は傑出した発明家であり、大事業家だった。言葉の原文は「There's a way to do it better – find it」。1957年からニューヨークタイムズなどでMcGraw-Edison Companyが展開した広告で、エジソンの絵像とともに頻繁に引用された。

編集後記

科学技術は軍事目的での研究開発によって大いに進歩してきた。最近の例を見ても、原子力、コンピューター、インターネット、GPSなど。これらが民生利用され、その便益を私たちも享受できるようになり、世の中が進歩したと感じる。

ロシアのウクライナ侵攻においても、ロシア側、ウクライナ側双方がAIを駆使しているといわれる。

これまでの科学技術は、人間がコントロールできたのだが、AIはインプットとアウトプットの間が人間の理解を超えてブラックボックス化し、またAI自体が自律的に発展して、人間のコントロールが効かなくなるのではという不安がつきまとう。

国際的に人間中心のAIの議論が進み、開発者や利用者（製品・サービスへ組み込む側）でも自主基準の策定も進むが、その思いを表紙の絵に込めた。

次号は「政策デザイン（仮）」を取り上げます。ご期待ください。　（編集長：花原克年）

Nextcom（ネクストコム） Vol.50 2022 Summer
2022年6月1日発行

発行　株式会社KDDI総合研究所
〒105-0001
東京都港区虎ノ門2-10-4オークラプレステージタワー
URL：www.kddi-research.jp

編集長　花原克年（株式会社KDDI総合研究所）
編集協力　株式会社ダイヤモンド社
株式会社メルプランニング
有限会社エクサピーコ（デザイン）
印刷　瞬報社写真印刷株式会社

株式会社 KDDI 総合研究所
http://www.kddi-research.jp
ISSN 2434-6233