2.1.4 AIソフトウェア工学

（1）研究開発領域の定義

AIソフトウェア工学は、AI（Artificial Intelligence：人工知能）応用システムを、その安全性・信頼性を確保しながら効率よく開発するための新世代のソフトウェア工学を指す[1)]。

従来型のシステム開発においては、安全性・信頼性を確保し、効率よくシステム開発を行うための技術体系・方法論がソフトウェア工学の中で整備されてきた。ここでいう従来型とは、プログラム（手続き）を書くという演繹型のシステム開発方法を意味する。これに対して、AI応用システムの開発では、データを例示することによる、機械学習を用いた帰納型の開発方法が用いられる。 AIソフトウェア工学は、従来の演繹型システム開発のためのソフトウェア工学から、AI応用システム向けの帰納型システム開発にも対応したソフトウェア工学へ拡張した技術体系・方法論である。

なお、AIソフトウェア工学とほぼ等しい用語として、国内では「機械学習工学」[2), 3), 4)]、海外では「Software 2.0」[5), 6)] がよく用いられている。

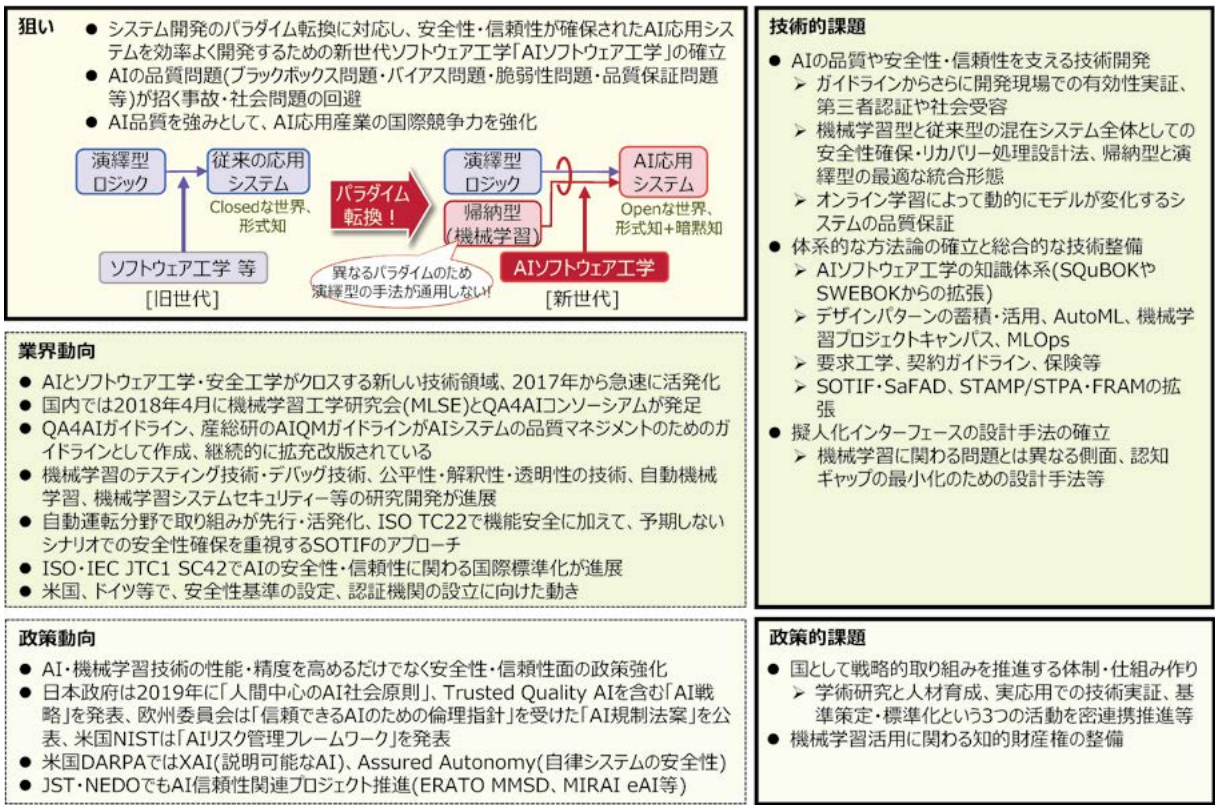

図2–1–6　　領域俯瞰：AIソフトウェア工学

（2）キーワード

機械学習工学、Software 2.0、AI応用システム、機械学習応用、AI品質、AI信頼性、AI公平性、説明可能AI、XAI、ブラックボックス問題、バイアス問題、ソフトウェアテスティング、訓練済みモデル、自動機械学習、AutoML、MLOps、SOTIF、SaFAD

（3）研究開発領域の概要

［本領域の意義］

現在のAIブームを牽引しているのは、深層学習（Deep Learning）をはじめとする機械学習技術の進化である。機械学習技術はさまざまな製品・システムに組み込まれ、実社会での応用・実用化が急速に広がっている。

しかし、機械学習による帰納型のシステム開発方法は、従来の開発スタイルとは大きく異なる。そのため、従来ソフトウェア工学として構築・整備されてきた技術・方法論（V字モデルなど）は必ずしも適さず、システムの要件定義や動作保証・品質保証にも新しい考え方が必要になる。システム開発のパラダイム転換が起きているのである[1), 2), 3), 4), 5), 6)]。

このパラダイム転換によって、システム開発に必要な人材スキルや方法論が刷新され、この変化に追従できないと、ソフトウェア産業やシステムインタグレーターは競争力を失いかねない。また、動作保証・品質保証などの考え方が整備されないまま、機械学習技術を組み込んだシステムが急激に社会に入っていくと、そこで発生した問題や事故が社会問題化する懸念もある。顕在化してきた問題として次のようなものが指摘されている[1]。

- ブラックボックス問題[7), 8), 9)]：判定理由について人間に理解可能な形で説明してくれない。事故発生時に原因解明や責任判断ができない。AIの解釈性・説明性が求められる。
- バイアス問題[10), 11), 12)]：訓練データ（学習データ）に偏見が含まれていると、判定結果に偏見が反映されてしまう。訓練データの分布の偏りが差別を生むこともある。AIの公平性が求められる。
- 脆弱性問題[28), 29)]：訓練（学習）範囲外のデータに対して、どう振る舞うかは不明である。敵対的サンプル（Adversarial Examples）[13)]と呼ばれる画像認識などの誤認識を誘発する攻撃[2]や、悪意を持った追加学習によって、不適切な振る舞いが引き起こされ得る。AIの安全性・頑健性が求められる。
- 品質保証問題：仕様（正動作）が定義されないため、テストの成否が定まらない。精度100%は無理で間違いは不可避で、動作保証が難しい。AIの信頼性が求められる。

「2.1.1 知覚・運動系のAI技術」「2.1.2 言語・知識系のAI技術」「2.1.3 エージェント技術」で述べてきたように、機械学習技術を用いることで、さまざまな応用において人間の判断を上回る精度の分類・予測・異常検知などが可能になった。これは、人間が形式知化できていないような規則性が機械学習技術によって獲得可能になり、コンピューターによってシステム化・自動化できる機能が広がってきたということである。 AIソフトウェア工学は、このような新たな価値を生み出すAI応用システムについて安全性・信頼性を確保するとともに、その効率の良い開発を可能にする。

［研究開発の動向］

❶ 学術界・産業界の動向

AIソフトウェア工学は、AIと、ソフトウェア工学（Software Engineering）や安全工学（Safety Engineering）がクロスする新しい技術領域である。国内では、2015年頃からシステム開発のパラダイム転換への対応が必要だという問題提起がされ始め、2017年初頭から学会・業界イベント（2017年2月の

1 これはAI応用システムの安全性・信頼性などが本質的に低いということを意味するわけではない。パラダイム転換に対して、システム開発のための技術体系・方法論がまだ整備されていないためである。 AIソフトウェア工学を確立していくことが、このような問題・懸念への対策になる。また、システムの安全性・信頼性の確保に向けては、開発者の視点だけでなく、システムの利用者や開発依頼者がAI応用システムの安全性・信頼性などに関する考え方や特性を理解し、どのように受け入れていくか、という側面も考えていく必要がある。

2 コンピューターセキュリティーインシデントに関する情報提供・技術支援を行っているJPCERTコーディネーションセンターから、脆弱性関連情報としてAdversarial Examplesに対する注意が発信された（2020年3月25日 JVNVU#99619336）。通常は特定製品に関する脆弱性が報告されるのに対して、アルゴリズムそのものに関する注意喚起が行われたのは異例のことである。

情報処理学会ソフトウェアジャパン2017、同年8月の情報処理学会ソフトウェア工学シンポジウムSES2017、他多数）で基調講演や企画セッションなどが立て続けに開催され、一気にホットトピック化した[1)]。

さらに2018年4月に、日本ソフトウェア科学会に機械学習工学研究会MLSE（Machine Learning Systems Engineering）が発足した[4), 14)]。 MLSEは、機械学習応用システムの開発・運用にまつわる生産性や品質の向上を追求する研究者とエンジニアが、互いの研究やプラクティスを共有し合う場として、研究発表会、ワークショップ、勉強会など、さまざまな活動を展開しており、この分野における日本の中核的コミュニティーになっている。機械学習を用いたシステムの要件定義から設計・開発・運用まで、プロセス管理や開発環境・ツール、テスト・品質保証の手法、プロジェクトマネジメントや組織論も含めて、機械学習を用いたシステム開発全般について幅広いスコープで活動している[4)]。

同じ2018年4月には、AIプロダクト品質保証コンソーシアムQA4AI（Consortium of Quality Assurance for Artificial–Intelligence–based products and services）も発足した[14)]。 AIプロダクトの品質保証に関する調査・体系化、適用支援・応用、研究開発を推進するとともに、AIプロダクトの品質に対する適切な理解を啓発する活動を行っている。2019年5月にはQA4AIの「AIプロダクト品質保証ガイドライン」（QA4AIガイドライン）[15)] が公開された。また、2020年6月には産業技術総合研究所から「機械学習品質マネジメントガイドライン」（AIQMガイドライン）[16)] が公開された。両ガイドラインとも継続的にアップデートされている。 QA4AIガイドラインは、品質保証で考慮すべき五つの軸が定義され、それぞれに関してチェックリストが示されている。また、生成系システム、Voice User Interface（スマートスピーカーなど）、産業用プロセス、自動運転、AI–OCR（機械学習を用いた光学文字認識）の5ドメインについて、個別のガイドラインも例示されている。一方、AIQMガイドラインは、利用時品質、外部品質、内部品質という三つを関係付けて、その向上のための要件を整理し、開発ライフサイクル全体としての品質管理の考え方を示している。相互に対応する部分は多く、二つのガイドラインは相補的な関係にある。いずれも産業界のメンバーを含めて検討・評価を行っており、産業界におけるAI品質管理の実践につながっている。これらのガイドラインは分野共通の基本的な考え方に重点が置かれており、産業界での実運用に際しては、分野ごとに具体化したガイドラインや事例集を用意するのが有効であり、そのような取り組みも行われている[3]。また、2022年には、AI品質管理に関わる法規・標準・ガイドラインの抽象的な要求事項を、AI品質管理の現場に迅速に適用し、AI品質を向上させる手法の開発を目的として、日本品質管理学会に「AI品質アジャイルガバナンス研究会」が発足した。

一方、海外でも2017年後半から、機械学習を従来型プログラミングに対する新しいパラダイムと捉える動きが見られた。新しいパラダイムは「Software 2.0」[5)] とも呼ばれ、国際学会（2018年6月のISCA 2018、2018年12月のNeurIPS 2018など）で基調講演[6)]も行われるようになった。カナダのモントリオール理工科大学にSEMLAイニシアチブ（The Software Engineering for Machine Learning Applications initiative）が発足し、2018年6月にキックオフシンポジウムが開催された。2018年9月にGoogle DeepMindが自社のAI開発ガイドラインを、仕様、頑健性、保証という3面からまとめたことを発表したのをはじめ、産業界でも自社のガイドラインを定める動きが国内外で広がっている。2021年7月にドイツのFraunhofer IAIS（Fraunhofer Institute for Intelligent Analysis and Information Systems）が公開したAI Assessment Catalog（Guideline for Designing Trustworthy Artificial Intelligence）は、日本のガイドライン（AIQM、QA4AI）と同様に詳細なガイドラインとして注目される。

3 例えば、石油・化学プラント向けガイドライン・事例集が策定されている。
https://www.meti.go.jp/press/2020/11/20201117001/20201117001.html（accessed 2023–02–01）

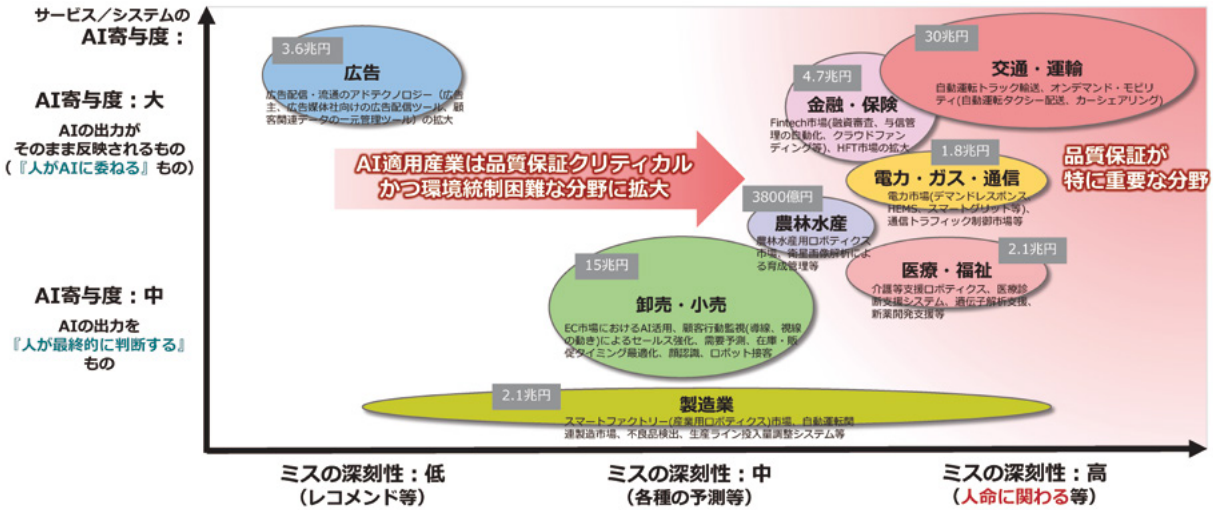

図2–1–7　産業分野と品質保証クリティカル性[4]

以上に示したような取り組みが活性化してきた背景には、AI・機械学習技術の応用がさまざまな分野に急速に広がり、品質保証クリティカルな応用分野にも適用されるようになってきたことがある（図2–1–7）。安全性・信頼性に関する要求レベルは応用ごとに異なる。応用システムが持つ三つの性質「ミスの深刻性」「AI寄与度」「環境統制困難性」[5]に着目すると、一般に、ミスの深刻性が高く、AI寄与度が高く、環境統制困難性の高い応用ほど、品質保証がクリティカルになる[1)]。機械学習の応用として、商品のレコメンド機能や、文字認識による郵便物の自動読み取り区分システムなどは2000年以前に実用化されているが、これらはミスの深刻性や環境統制困難性が比較的低い応用である。これに対して、昨今注目される自動運転や医療診断といった応用分野は、ミスの深刻性や環境統制困難性が高い（ミスが人命に関わり、多様な環境条件で使われる）。そのため、事前（および運用時）の品質保証が極めて重要なものになっている。

産業界の中でも特に問題意識が高く、検討が先行しているのが自動車業界である。自動運転の実現に向けてAI・機械学習技術の役割が増しており、上でも述べたように品質保証クリティカルな応用分野として具体的な検討が進められている[17), 35)]。国際的には、ISO TC22などで自動運転の安全性規格の策定が進められている（詳細は［注目すべき国内外のプロジェクト］❶を参照）。国内でもデンソーが自動運転を含むAI搭載システムの品質保証のための仕組み作りや技術開発[36)]を進めているなど、取り組みが活発化している。

また、産業界では、品質管理・安全性確保という面にとどまらず、機械学習応用システム開発・運用のプロセス全体にわたって効率化・最適化していくため、新しい考え方・フレームワークが検討されている[18)]。従来、開発側（Dev）と運用側（Ops）が協調した取り組み・フレームワークとしてDevOpsがあるが、これを機械学習（ML）応用システムの開発・運用に発展させたものがMLOpsと呼ばれている。機械学習

4　文献1）から再掲。図中の金額は2030年のAI適用産業の予想市場規模であり、EY総合研究所のレポート「人工知能が経営にもたらす創造と破壊」（2015年9月）をもとにした。
https://kyodonewsprwire.jp/prwfile/release/M103415/201509143541/_prw_OA1fl_O8ov31l1.pdf（accessed 2023–02–01）

5　「ミスの深刻性」は、AI・機械学習が誤った判定結果を出したときに生じる問題がどれくらい深刻であるかを意味する。人命に関わるような場合は深刻性が高い。「AI寄与度」は、問題解決のために実行されるアクションの決定にAI・機械学習がどれくらい大きく寄与するかを意味する。AI・機械学習の出力（判定結果）がそのまま反映される場合は寄与度が高く、AI・機械学習の出力（判定結果）を参考にして人間が最終的に判断する場合は寄与度が低い。「環境統制困難性」は、AI・機械学習を実行する際の環境条件をコントロールすることの難しさを意味する。環境条件を列挙することが難しく想定外のことがいろいろ起こり得る場合は困難性が高く、環境条件を統制することが容易であれば困難性が低い。

型プロジェクトにおける要検討事項を整理した「機械学習プロジェクトキャンパス」[6]も活用されている。技術的には、さまざまな機械学習のアルゴリズムやそのパタメーターから、より良いものを選び、機械学習を用いた分析を自動化する自動機械学習（AutoML）も実用化されている。DataRobotや、NECからはカーブアウトしたdotDataをはじめ、Google Cloud AutoML、Microsoft Azure Machine Learning、H2O Driverless AI他、各社からAutoMLサービスが提供されている。また、設計ノウハウをデザインパターンとしてカタログ化して蓄積・活用する取り組みも進んでいる[19)]。データに基づいてソフトウェアの生産性や信頼性の向上を図る実証的ソフトウェア工学のアプローチを、機械学習システムに適用する取り組み（例えばバグの実態調査・分析など）も行われている。

その一方、AI・機械学習の研究者は、機械学習の精度・性能を高める競争が激しく、総じて開発法自体への関心は低かったが、この4–5年で社会におけるAIについての議論が活発に行われるようになり、機械学習の脆弱性問題・バイアス問題・ブラックボックス問題などへの対処を中心に取り組まれるようになった。2014年からFAT/ML（Fairness, Accountability, and Transparency in Machine Learning）ワークショップ、2018年からはACM FAT*が開催されているほか、AIの主要国際会議（AAAI・NeurIPSなど）でも研究発表が増えている。また、2019年12月には、国内の機械学習の研究者コミュニティー（人工知能学会倫理委員会、日本ソフトウェア科学会機械学習工学研究会、電子情報通信学会情報論的学習理論と機械学習研究会）が共同で「機械学習と公平性に関する声明」を発表し、翌月それを受けたシンポジウムを開催した。

❷ 基準策定・標準化の動向

上述のような問題意識の急速な高まりと連動して、AI品質関連の標準化活動や安全基準策定活動が多数進められている。AIに関する主な標準化委員会としては以下が挙げられる[20)]。

国際標準化機構ISO（International Organization for Standardization）と国際電気標準会議IEC（International Electrotechnical Commission）の第1合同技術委員会JTC1において、SC7がソフトウェア工学、SC42が人工知能を扱っており（JTC：Joint Technical Committee、SC：Subcommittee）、AIの品質や安全性・信頼性に関わる議論が進められている。特に2017年に立ち上がったSC42では、基盤的規格群に関するWG1、ビッグデータに関するWG2、Trustworthinessに関するWG3、ユースケースに関するWG4、計算的アプローチと特性に関するWG5、AIのガバナンスに関するSC40との合同WG、AIベースシステムのテスト法に関するSC7との合同WGが活動している（WG：Working Group）。これに対応するための国内のミラー委員会として、情報処理学会技術規格調査会に「SC42専門委員会」が設置されて活動している。前述のAIQMガイドラインもSC42の活動に反映されている。

また、米国電気電子工学会（IEEE）では、2019年3月に「倫理的に配慮されたデザイン（Ethically Aligned Design）」というレポートの第1版（EAD1e）を公表したが（詳細は「2.1.9 社会におけるAI」を参照）、これと連動する標準化プロジェクトとしてIEEE Standard Association（IEEE–SA）のP7000～P7014が進められている。 IEEE–SAの提案により、AIシステム開発のための標準規格に関心を持つ機関に議論や協調の場を与える国際フォーラムOCEANIS（The Open Community for Ethics in Autonomous and Intelligent Systems）も設立された。他にもIECで標準管理評議会（SMB）システム評価グループ（SEG）10「Ethics in Autonomous and AI Applications」が、AIアプリケーションの倫理的側面に関する（IEC委員会に広く適用される）ガイドラインの作成などを目的に設置された。

6 ビジネスモデル構築における要検討事項を整理したものとしてよく知られた「ビジネスモデルキャンバス」を参考にしつつ、機械学習型プロジェクト向けに12の要検討事項が整理されている。三菱ケミカルホールディングスが自社の経験に基づいて体系的に整理し、2019年7月に一般に公開した。

さらに、欧州委員会は2021年4月に「AI規制法案（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）」を公表した。この法案では、AI応用システムをリスクの大きさに着目して四つのレベルに分け、そのレベルに応じて使用禁止や適合性評価の義務化など、かなり踏み込んだ規制（ハードロー）をかけようとしている。意見公募を経て修正案が検討されており、2023年中の発効が有力で、完全施行は最速で2024年後半と見られている。また、米国では、2023年1月に標準技術研究所（National Institute of Standards and Technology：NIST）から「AIリスク管理フレームワーク（Artificial Intelligence Risk Management Framework：AI RMF）」が発表された。 AIのリスクに対する考え方（技術属性、社会技術属性、信頼原則に分けて考えるなど）やリスクに対処するための実務が示されている。これら欧州AI規制法案と米国NIST AI RMF（「2.1.9 社会におけるAI」でも記載している）は、国際標準化への影響も見込まれる。

ドイツでは、ドイツ人工知能研究センター（The German Research Center for Artificial Intelligence：DFKI）とドイツの認証機関TÜV SÜDが共同で、TÜV for Artificial Intelligence策定の活動を進めている。日本国内では、ソフトウェア品質知識体系SQuBOK（Software Quality Body of Knowledge）が、2020年11月にV3に改訂され、新たにAI応用システムの品質に関わる内容が追加された[21)]。

なお、特定業界の安全規格の策定は、自動運転分野が特に進んでいるが、その状況については［注目すべき国内外のプロジェクト］❶に記載する。

❸ 科学技術政策の動向

AIに関する科学技術政策は、いま各国が国としての戦略を掲げ、重点投資を進めている。その中で、AI・機械学習技術の性能・精度を高める技術開発競争が強く意識されてきたが、徐々に安全性・信頼性の面にも目が向けられるようになってきた。

わが国では、総理指示を受けたAI研究の体制として、2016年に「人工知能技術戦略会議」とその下での総務省・文部科学省・経済産業省の3省連携による推進体制が構築された。さらに、2018年6月の閣議決定を受けて「統合イノベーション戦略推進会議」が設置され、2019年3月に「人間中心のAI社会原則」、2019年6月に「AI戦略2019」が決定された。「人間中心のAI社会原則」では、AIの社会的・倫理的・法的な課題（Ethical, Legal and Social Issues：ELSI）を含む社会から見たAIへの要請[7]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）セキュリティー確保の原則、（5）公正競争確保の原則、（6）公平性、説明責任および透明性の原則、（7）イノベーションの原則という七つを掲げた。「AI戦略2019」ではAI社会原則を受けて、「Trusted Quality AI」[8]がAI研究開発の中核的課題として位置付けられた。 AI原則を満たす「信頼される高品質なAI」を実現するための技術開発、すなわちAIソフトウェア工学の必要性が認識され、国立研究開発法人科学技術振興機構（JST）や国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の研究開発プログラムが

7 社会から見たAIへの要請については、「人間中心のAI社会原則検討会議」（2018年4月発足）に先立ち、内閣府の「人工知能と人間社会に関する懇談会」、総務省の「AIネットワーク社会推進会議」、経済産業省の「AI・データ契約ガイドライン検討会」などで検討されてきており、それらを踏まえて「人間中心のAI社会原則」が検討された。特にAIネットワーク社会推進会議は、2017年7月に、連携、透明性、制御可能性、安全、セキュリティー、プライバシー、倫理、利用者支援、アカウンタビリティーという9つのAI開発原則を掲げた「国際的な議論のためのAI開発ガイドライン案」を公表し、2019年8月には「AI利活用ガイドライン」も公表した。

8 「AI戦略2019」に先立ち2019年2月に一般社団法人日本経済団体連合会（経団連）から発表された「AI活用戦略」にも「Trusted Quality AI」のコンセプトが示されている。 AI戦略はその後「AI戦略2021」「AI戦略2022」とアップデートされたが、Trusted Quality AIは重要な研究開発課題として位置付けられている。

推進されている（［注目すべき国内外のプロジェクト］❷を参照）。

AI社会原則・AI倫理指針は、2019年にG20（主要20カ国・地域首脳会議）で取り上げられ、各国から同様のものが次々に発表された。欧州委員会の「信頼できるAIのための倫理指針」（Ethics Guidelines for trustworthy AI、2019年4月）、OECD（経済協力開発機構）の「人工知能に関するOECD原則」（OECD Principles on Artificial Intelligence、2019年5月、42カ国署名）、ユネスコ（国際連合教育科学文化機関）の「AI倫理勧告」（first draft of the Recommendation on the Ethics of Artificial Intelligence、2021年11月、全193加盟国採択）などが挙げられる（詳細および関連動向は「2.1.9 社会におけるAI」参照）。関連する具体的な研究開発投資では、米国の国防高等研究計画局（Defense Advanced Research Projects Agency：DARPA）が推進するXAI（Explainable AI）プロジェクト（2017年5月～2021年4月）とAssured Autonomyプロジェクト（2018年5月～2022年4月）がAIの安全性・信頼性に関するものとして知られている。 XAIプロジェクトでは、人間の意思決定を支援するパートナーとしてのAIを、人間の兵士が理解・信頼し、管理することを目指し、具体的な目標として、マルチメディアデータからターゲットを選択する際の理由説明を扱うData Analyticsタスクと、ドローンやロボットなどの自律システムがどういう状況でどういう理由で次の行動を決定したかを説明するAutonomyタスクが設定された。Assured Autonomyプロジェクトでは、自動運転車やドローンなどの自律システムの安全性確保が研究された。

（4）注目動向

［新展開・技術トピックス］

❶ 機械学習のテストおよびデバッグの技術

機械学習を用いて作られたシステムは、どれだけテストすれば十分なのか、テストの方法や品質指標がまだ確立されていない。従来型の簡単なテストのイメージは、「ボタンAを押したら光る」「ボタンBを押したら音が鳴る」というような動作ロジックに沿って、すべてのケースと正しい結果を事前にリストアップすることができ、その通りの結果が得られるかを確認すればよいというものである。従来型でも、一定以上の規模や複雑さを持つシステムになると、すべてのケースはリストアップできず、テストが難しくなるが、機械学習型の場合は、動作ロジックの記述ではなくデータ例示によって動作を定義するので、そもそもすべてのケースをリストアップするための手掛かりがない。例えば、自動運転における環境認識では、「雨や霧のこともあるかもしれない」「物陰から人が飛び出すかもしれない」など、実世界で車が遭遇し得る環境の可能性をどう数え上げ、どれだけのケースをテストしておいたら十分安全なシステムだと言えるのか、という問題は機械学習型においていっそう難しい。

このことから、事前に想定していなかったケースに対するシステムの振る舞いが保証できず、脆弱性が生じる。この脆弱性を突く攻撃がAdversarial Examples攻撃[13]である。例えば、機械学習を用いた画像認識システムがそれまで正しく認識できていた画像に対して、人間には気にならない程度の小さな加工（ごく小さなノイズなど）を加えて、それまでと全く異なる誤認識結果を出させるというものである。道路標識を対象とした実験で、停止標識を速度制限標識と誤認識させた（停止しなければ事故を招く）と報告されている。

機械学習のテスティング技術は、AIソフトウェア工学分野で特に活発に取り組まれている研究テーマの一つである[22]。機械学習は訓練データによってシステムの動作が定まるが、起こり得るすべてのケースを訓練データやテストデータとして事前にカバーすることはできない。そのような前提のもと、テストデータのカバレッジやパターン量を適切かつ効率よく増やすためのさまざまな手法が開発されている[4), 22), 23)]。具体的には、ニューロンカバレッジ、メタモルフィックテスティング、サーチベースドテスティング、データセット多様性などのアイデアが知られている。ニューロンカバレッジは、深層学習などニューラルネットワーク系の機械学習において、ニューラルネットワーク内の活性化範囲を調べ、それを広げるようなテストパターンを

生成しようという考え方である。メタモルフィックテスティングは、入力を変えると出力はこう変わるはずという関係を検証し、既存テストケースから多数のテストケースを生成する手法である。サーチベースドテスティングは、メタヒューリスティックを用いて、欲しいテストケースを表すスコアを最大化するようテストケースを生成する手法である。また、適用できるケースはまだ限定的だが、形式検証を深層学習のモデルにも適用する試みも進みつつあり、そのコンペティション（VNN–COMP）も開催されている。個別失敗ケースを見つけるのではなく、弱点領域・性能限界を追求するという取り組みもあり、機械学習のテスティング技術の研究は多様化している。その一方で、開発現場で十分な意義・効果が示されているか、費用対効果が妥当かなどは、まだ十分に見極められていない。

また、時間の経過とともに、入力されるデータの傾向や形式が変化したり、結果を利用する側の評価基準が変化したりして、実質的な精度が低下してしまうことがよく起こることから、機械学習システムでは、導入前の評価・テストだけでなく、運用中の評価・テストも重要である。入力データや出力結果をモニタリング・評価して、その変化が許容できるレベルを超えたら、調整・再学習するような仕掛けも用いられている。

また、問題を見つけるだけでなく原因把握・問題箇所特定・デバッグの手法も検討されるようになった[24), 25)]。成功ケースと失敗ケースの比較からニューラルネットワークの再訓練箇所を決める方法、再訓練で追加すべきデータを選択する方法、再訓練ではなくパラメーターを直接修正する方法などが提案されている。機械学習システムは、ある問題に対処するために加えた部分的な修正が、そのシステム全体の動作に影響を与えてしまうというCACE性（Changing anything changes everything）があり、これがデバッグを難しいものにしているが、影響を与える範囲を絞りつつ、問題に効果的に対処するような修正を求める方法も追求されている。

❷ 機械学習における公平性・解釈性・透明性（FAT/ML）

上記❶で述べたようなさまざまなケースをテストすることに加えて、FATと呼ばれる公平性・解釈性・透明性を、機械学習の応用システムにおいてどう確保するか、というのも重要課題である。特に解釈性と公平性に対する技術的対策がホットトピックになっている。透明性に関しては、解釈性・公平性と併せて実装されるが、欧州で2018年5月に施行された一般データ保護規制（General Data Protection Regulation：GDPR）にAIの透明性を求める条文（GDPR第22条）が盛り込まれたことから、規制順守という面での対応も求められる。

公平性の確保、すなわちバイアス問題への対策では、公平性配慮データマイニング（Fairness–aware Data Mining：FADM）[12)] あるいは公平性配慮機械学習（Fairness–aware Machine Learning）呼ばれる研究トピックが立ち上がっている。機械学習におけるバイアス問題は、主に訓練データの分布の偏りや正解ラベル付けへの偏見混入などによって、人種・性別のようなセンシティブ属性が判定結果に大きく関わることで起きる。例えば、特定人種の誤認識が多いとか、性別によって採用判定が左右されるとかの不公平な結果が生じる。人種・性別の属性を除外して機械学習にかけたとしても、他の属性に人種・性別と相関の高いものがあれば、不公平な結果になり得る。 FADMでは、グループ公平性・個人公平性などの公平性基準を定義し、それを用いて不公平さを検出する手法や、不公平を防止する手法が提案されている。そのためのツールとして、MicrosoftのFairlearn、IBMのAI Fairness 360、GoogleのFairness Indicatorなども提供されている[9]。ただし、公平性を確保することで通常、精度は低下することや、どの公平性基準を採用するかによって結果が異なることなどを理解し、応用ごとの要件を明確化して設計すること

9 このようなツールでよく活用されている公平性指標として、グループ公平性に基づくDemographic ParityやEqualized oddsなどがある。例えば、Demographic Parityでは「男女で採用率が同じ」、Equalized oddsでは「男女で判断基準が同じ」のようなことを意図した指標である。

が必要である。

解釈性の向上、すなわちブラックボックス問題への対策は、XAI（Explainable AI：説明可能AI）[7), 8), 9)]と呼ばれ、活発に取り組まれている研究トピックである。研究論文数が急増するとともに、OSS（Open Source Software）や商用ソフトウェアとして開発現場での活用も広がっている。 XAI技術を大きく分けると、（A）深層学習のように精度が高いが解釈性が低いブラックボックス型の解釈性を高めるアプローチと、（B）決定木や線形回帰のような解釈性は高くても精度に限界のあったホワイトボックス型の精度を高めるアプローチがある。アプローチ（A）には複数のタイプがあり、その一つは（A–1）大域的な説明と呼ばれるもので、ブラックボックス型を近似するようなホワイトボックス型モデルを外付けするという方法である。例えば、深層学習の結果を決定木で近似するような試みがある（Born Again Treesなど）。もう一つは（A–2）局所的な説明と呼ばれるもので、ブラックボックス型がある判定結果を出したときに、その結果が出た要因を示すという方法である。 LIME、Influence、Attention Mapなどのツールがよく知られている。さらに（A–3）人間の知見を埋め込むという方法がある。人間が定めた分類の着眼点が分かっているときなど、人間の知見をモデルの制約として与えつつ学習させるというものである（例えばAttention Branch Network）。一方、アプローチ（B）の例としては、決定木的な場合分けと重回帰分析式を合わせて最適化する異種混合学習技術（因子化漸近ベイズ推論）が実用化されている。解釈性と精度もトレードオフ関係にあり、説明の目的も、製品・システムの品質保証、事故・問題発生時の説明責任、開発の効率化（デバッグ）、ユーザーから見た安心感・信頼の確保など、さまざまであるから、そこで求められる解釈性の要件に応じて適切な方法を選択することが必要である。また、解釈・説明は近似的なものになるので、だますような説明（Fairness Washing）も作れてしまう可能性がある[26)]。そういった点も考慮し、XAIの評価についても検討されている。これに関連して、米国の国立標準技術研究所（National Institute of Standards and Technology：NIST）は、2020年8月に「Four Principles of Explainable Artificial Intelligence」と題したドラフトレポートを公開した[27)]。このレポートでは、XAIの4原則として、Explanation（結果に対するエビデンスや理由を示すべき）、Meaningful（ユーザーに理解可能な説明を提供すべき）、Explanation Accuracy（説明は結果を出すプロセスを正確に反映すべき）、Knowledge Limits（システムは設計された条件下か、結果に十分な確信があるときのみ動作する）を挙げている。

❸ 機械学習システムセキュリティー

機械学習システムには、それ特有の脅威がある。これを突く攻撃の種類として、以下のようなものがよく知られている[28), 29)]。

まず、誤認識や想定外動作を誘発する攻撃として、回避攻撃（Evasion Attack）やポイズニング攻撃（Poisoning Attack）などがある。回避攻撃は、機械学習システムへの入力に、悪意のある変更を加えることで、システムに誤動作をさせる。入力データに人間には気にならない程度のノイズを加えることで、モデルの誤判断を誘発する敵対的サンプル（Adversarial Example）と呼ばれる攻撃がよく知られている。ポイズニング攻撃は、訓練データやモデルに細工をすることで、システムに誤動作をさせる。データやモデルをある程度汚染して認識性能を低下させる攻撃のほか、特定データをトリガーとして誤動作を引き起こすように仕込むバックドア攻撃が知られている。

また、モデルやデータを窃取する攻撃として、モデル抽出攻撃（Model Extraction Attack）、モデルインバージョン攻撃（Model Inversion Attack）、メンバーシップ推測攻撃（Membership Inference Attack）などがある。これらは、対象とする機械学習モデルの入力と出力の関係を手掛かりとして、モデル抽出攻撃は同等の性能を持つモデルを作成し、モデルインバージョン攻撃は訓練データに含まれる情報を復元し、メンバーシップ推測攻撃はあるデータが訓練データ中に含まれるかを特定するものである。

このような攻撃への対策としては、モデルの頑健性を評価し、訓練方法を改良することや、差分プライバシー技術（「2.4.3 データ・コンテンツのセキュリティー」参照）や暗号化データ処理を用いて、データ内

容を読み取り困難にすることや、攻撃を検知して対策を起動するなどのシステムレベルの防衛などが考えられている。

このような研究に基づき、機械学習システムの開発過程において、上記のような攻撃のリスク分析（影響分析や脅威分析）とその対策（攻撃の検知と対処）をどのように進めるべきかについて、ガイドラインの形で体系化が進められている。その代表的なものとして、機械学習工学研究会MLSEの「機械学習システムセキュリティガイドライン」[29] や、前述のAIQMガイドラインが挙げられる。

また、内閣府が主導する経済安全保障重要技術育成プログラム（K Program）において、個別研究型の研究開発構想の一つとして「人工知能（AI）が浸透するデータ駆動型の経済社会に必要なAIセキュリティ技術の確立」が挙げられており、ここで述べたような研究開発課題への取り組みが強化されつつある。

［注目すべき国内外のプロジェクト］

❶ 自動運転のAI安全性に関するプロジェクト

自動運転の安全性評価に関わるプロジェクトとして、2016年1月～2019年6月に、ドイツ経済エネルギー省（Bundesministerium für Wirtschaft und Energie：BMWi）が主導して実施されたペガサスプロジェクトがある。しかし、ここで検討されたのは自動運転システム全体としてのシナリオベース検証であり、その中で使われるAI技術そのものの安全性要件は明に論じられていない。日本においても、戦略的イノベーション創造プログラム（SIP）で「自動運転」（SIP–adus）が推進されており、同様に自動運転システム全体としての安全性確保のための取り組みが進んでいる。

AI技術が取り上げられたものとして、2019年6月にSaFAD（Safety First for Automated Driving）ホワイトペーパー[30] が公開された。 Aptiv、Audi、Baidu、BMW、Continental、Fiat Chrysler Automobiles、Daimler、HERE、Infineon、Intel、Volkswagenの11社によるコンソーシアムが、安全性を考慮した自動運転システムを開発するための技術や検討事項について指針をまとめたものである。この中では、自動運転における機械学習ベースの画像認識システムを開発するときのプロセス・成果物・技術課題も取り上げられた。2020年には、SaFADをほぼ踏襲し、今後の自動運転安全性に関するISO標準化の基礎とする目的でまとめられた技術報告ISO TR 4804：2020が発行され、その後継となる技術標準ISO TS 5083（自動運転システムの安全）が検討されている。

自動車業界の国際標準はISO TC22（自動車）で作られる。従来適用されているのは、故障時のリスクを回避・低減する機能安全の規格ISO 26262である。しかし、AI応用では、誤認識や未学習ケースなど、故障以外の要因でリスクが多々発生する。そこで、2022年6月に、新たにISO 21448:2022（SOTIF）が発行された。SOTIF（Safety of the Intended Functionality）は、予期しないシナリオが発生したときの安全性確保を重視し、懸念されるケース・条件での動作が適切かどうかをひとつひとつ検討するアプローチをとる。未知のシナリオと危害が及ぶシナリオを特定して検証する作業を反復的に実行し、それを既知かつ危害のないシナリオに変えていくことで安全性を確保する。さらに、AIの安全性に関する公開仕様書ISO PAS 8800が2023年10月頃の発行を目指して議論されている。

なお、欧州では、前述の通りAI規制法案が定められつつあるが、厳格な既存規制が存在する応用分野では、その既存規制にAI要件を追加する形での運用が優先される見込みである。自動車業界については、車両型式認証がその既存規制に該当し、これにAI要件を加えていくことで、AI規制法案から除外されると考えられている。また、国連欧州経済委員会（United Nations Economic Commission for Europe）の自動車基準調和世界フォーラム（WP29）配下の自動運転技術分科会GRVA（Working Group on Automated/Autonomous and Connected Vehicles）からAIガイダンス案が発行され、それをたたき台としてAI用語定義などの議論が行われている。

米国では、第三者認証機関であるUnderwriters Laboratoriesが2020年4月にUL4600を発表した。これはドライバーが操作しない状態を主とする自動運転レベル4以上を想定して、安全性評価のための原則

とプロセスを示している。特定技術の使用は義務付けておらず、設計プロセスの柔軟性を許容している。また、安全性に関する合格/不合格といった基準や、実走行試験や倫理的側面の要件も定めておらず、急速に進歩する技術を過度に制約せずに安全性を確保する柔軟な規格としている。

また、AI安全性に関する研究プロジェクトとしても、自動運転をターゲットあるいは具体的検討事例としたものが多く見られる（米国カリフォルニア大学バークレー校のVerifAI、英国ヨーク大学のSafe Autonomy、カナダのウォータールー大学のWiSE Drive、日本ではERATO MMSDなど）。自動車業界と学術界の両面で活発に取り組まれている。

❷ JST・NEDOのAI信頼性関連のプロジェクト

国内では［研究開発の動向］❸で述べたように、国のファンドによるプロジェクトが、JSTやNEDOの研究開発プログラムとして推進されている。NEDOでは「次世代人工知能・ロボット中核技術開発」プロジェクト（2015年度～、プロジェクトマネージャー：渡邊恒文、プロジェクトリーダー：辻井潤一）において、2019年度に「人工知能の信頼性に関する技術開発事業」が実施され、「説明できるAI」で7件、「AI品質」で1件が採択された。これを引き継いで「人と共に進化する次世代人工知能に関する技術開発事業」（2020年度～2024年度）が立ち上がり、2020年度に19件のテーマが採択・実施されているが、そのうち6件が「説明できるAIの基盤技術開発」、1件が「実世界で信頼できるAIの評価・管理手法の確立」[10]に関するテーマである。

JSTでは、「ERATO蓮尾メタ数理システムデザイン（MMSD）プロジェクト」（2016年10月～2025年3月[11]、研究総括：蓮尾一郎）が、数理的基盤・形式手法などを活用して、物理情報システムや機械学習システムのような不確かさを内包する情報システムの安全性を保証する技術開発にチャレンジしている。自動運転システムの安全性保障が具体的ターゲットに設定されており、国際規格化（IEEE P2846）が議論されている責任感知型安全論（Responsibility–Sensitive Safety：RSS）を拡張することで、現実の複雑な運転シナリオに対しても、数学的な裏付けを持って、目標達成と安全性の両方を満たす枠組みGA–RSS（Goal–Aware RSS）[31)]の開発などが進められている。

また、JST未来社会創造事業「超スマート社会の実現」領域で「機械学習を用いたシステムの高品質化・実用化を加速する“Engineerable AI”技術の開発」（2020年4月～、研究開発代表者：石川冬樹、略称：eAIプロジェクト）[12]が本格研究として採択され推進されている。医療や自動運転など安全性・信頼性が重要となる応用分野を重点ターゲットとし、細やかなニーズに応えるAIシステムのためのeAI技術を研究開発している。特に、データが少ない状況でも安全性などの観点で重要なケースに対応したり、前述したCACE性の問題に対して修正の影響範囲を少なく抑えたりする技術開発を進めている。

さらに、文部科学省の2020年度戦略目標の一つとして「信頼されるAI」が定められ、それを受けたJST CREST「信頼されるAIシステムを支える基盤技術」（研究総括：相澤彰子）、JSTさきがけ「信頼されるAIの基盤技術」（研究総括：有村博紀）も実施されている。

10 産業技術総合研究所の機械学習品質マネジメントガイドラインや機械学習品質管理テストベッドの研究開発は、NEDOの人工知能の信頼性に関する技術開発事業、人と共に進化する次世代人工知能に関する技術開発事業からファンドを受けている。

11 当初は2022年3月までのプロジェクトだったが、追加支援期間（機関継承型）の枠組みにより期間が3年間延長された。

12 2020年4月～12月の探索研究期間を経て、2021年1月より本格研究に移行した。また、このプロジェクトの前身として、探索研究「高信頼な機械学習応用システムによる価値創造」（2018年11月～2020年3月、研究開発代表者：吉岡信和）が行われた。

（5）科学技術的課題

❶ AIの品質や安全性・信頼性を支える技術開発

AI・機械学習の品質や応用システムの安全性・信頼性を確保するための技術開発は、いっそうの強化が望まれる。二つの品質管理ガイドライン（QA4AIとAIQM）がまとめられ、年々拡充されているが、実際に開発の現場での活用に十分か、あるいは、AI品質に関する第三者認証機関のような形で運用するのに十分か、AI品質に関しての社会受容を得るのに十分かなど、実践しながらさらに内容が整備・拡充されていくことが期待される。

また、機械学習型コンポーネントは100%保証ができないものであることや、ブラックボックス型機械学習モデルの解釈性・説明性はあくまで近似的なものであることを踏まえると、機械学習型コンポーネント単体での保証は限界がある。従来型と機械学習型の混在システム全体としての安全性評価法やリカバリー処理設計法が必要である。例えばSafe LearningやSafety Envelope[32]のような従来型（演繹型）で安全性を確保した範囲内で機械学習（帰納型）を使う設計法や、機械学習型コンポーネントの入力・出力をモニタリングして例外処理・リカバリー処理を起動するシステム構成法が検討されている。AIの基本アーキテクチャー自体が、機械学習のような帰納型だけでなく、知識・記号推論のような演繹型と融合させた次世代AIアーキテクチャー[33]へと発展しつつあるので、そのような面からも帰納型・演繹型の最適な統合形態とその安全性・信頼性確保を考えていくべきであろう。

さらに、今後のAIシステムの発展を考えるならば、オンライン学習によって動的にモデルが変化するシステムの品質保証も大きな課題となる。機械学習は、訓練データを与えてモデルを生成する訓練フェーズ（学習フェーズとも呼ばれる）と、その訓練済みモデルを用いて、新たに入力されたデータを判定する判定フェーズ（推論フェーズや予測フェーズとも呼ばれる）を持つ。これまで検討されてきた品質保証法は基本的に、訓練フェーズのバッチ的実行を想定している。すなわち、初期の訓練であれ、追加の訓練であれ、訓練フェーズを実行したら、判定フェーズに入る前に、必ず訓練済みモデルを評価・テストがされなければならない。しかし、機械学習の使い方として、オンライン学習によって、モデルを随時更新しながら、判定にも使っていく形があり得る。このような形の場合、システムの品質保証ははるかに難しく、新たな技術チャレンジが必要である。

❷ 体系的な方法論の確立と総合的な技術整備

システムの安全性・信頼性の確保や新たなパラダイムでの開発効率化は、一つの技術で解決・達成できるものではなく、体系的な方法論の確立とそこで必要になる技術のバランスの良い整備を進めていく必要がある。その際、開発・運用プロセスの全体像を押さえつつ、必要な技術群を多面的・総合的に整備していくべきであろう[1]。SQuBOKやSWEBOKのようなソフトウェア工学の知識体系は参考になるであろうし、前述したように、デザインパターンの蓄積・活用や自動機械学習（AutoML）の活用による開発効率化、機械学習プロジェクトキャンパスやMLOpsの実践も進められている。また、機械学習ベースの帰納的開発では、従来の演繹型開発のような動作仕様が定められないことや、性能保証ができないことなど、要求分析時の不確実性が大きい[4]。そのため、開発にかかる工数・費用の見積もりが難しく、開発完了に関わる出荷判定や検収条件でも問題が生じやすい。要求工学、契約ガイドライン[13]、保険などの面からも検討・整備が必要になっている。

また、従来のソフトウェア工学との対比で語られることが多いが、AI・機械学習の応用システム開発は、ソフトウェアだけに閉じず、ハードウェアやデータ管理も含めたシステムとして考える必要がある。狭い意味

13 経済産業省による「AI・データの利用に関する契約ガイドライン」、日本ディープラーニング協会による「契約締結におけるAI品質ハンドブック」が策定され公開されている。

でのソフトウェア工学に限らず、安全工学やシステム工学も検討範囲に含まれる。例えば、エッジケースに着目したSOTIF・SaFADや、機能間の関係性を踏まえて制御系が環境と相互作用することで起きうる事故モデルを使った安全性分析・ハザード解析手法として知られるSTAMP/STPA[34]（System Theoretic Accident Model and Processes / System Theoretic Process. Analysis）やFRAM（Functional Resonance Analysis Method）などのAI・機械学習の応用システムへの拡張適用[14]なども検討されている。このような多面的な取り組みを進めつつ、それらを体系的な方法論、総合的な技術群として整備していくことが望まれる。

さらに、AIシステムの安全性・信頼性などの品質マネジメントを含みつつ、より広くAIのELSIについて、原則レベルから実践フェーズへの移行が進み、AIガバナンスとしての取り組みも重要視されるようになってきた。AIガバナンスの考え方やフレームワークについては「2.1.9 社会におけるAI」で取り上げる。

❸ 擬人化インターフェース設計に関する方法論・技術

ここまで、AI応用システム開発における問題を、機械学習に起因するものにフォーカスして論じたが、機械学習以外にも問題になり得る要因が考えられる。その一つは擬人化インターフェースである。2次元（画面表示）にせよ3次元（ロボット形状）にせよ、人間の形状・表情・対話を模したインターフェース（擬人化インターフェース）を持つAI応用システムが提供されつつある。擬人化インターフェースの利点は、そのシステムと相対する利用者にとって、システムがどのような応答をするかのモデルを仮定しやすいことである。しかし、それは逆に、利用者が思い込みをしやすい面があり、利用者が仮定したモデルと、実際のシステムの応答モデルとの間のギャップが、想定外の状況を生む可能性を持つ。これはヒューマンエージェントインタラクション（HAI）の研究において「適応ギャップ」と呼ばれる問題である（「2.1.4 エージェント技術」を参照）。この適応ギャップを最小化するような設計手法が求められる。

❹ 生成AIを用いた開発手法

「2.1.2 言語・知識系のAI技術」で述べたように、生成AI技術をソフトウェア開発に活用することが行われつつある。 ChatGPTやText–to–Imageなどのテキスト・対話からの生成系AIをコード生成に応用できるほか、DeepMindのAlphaCodeやOpenAI Codexなどプログラムコード生成向けに事前学習したツールも提供されている。このような技術を使って開発効率を高める方法論や品質確保手法も今後の研究開発課題である。

（6）その他の課題

❶ 国として戦略的取り組みを推進する体制・仕組み作り

［研究開発の動向］❸や［注目すべき国内外のプロジェクト］❷で述べたように、国のAI戦略の中でも言及され、NEDOやJSTのプログラムが実施されている。産業界で活用できるガイドライン（QA4AIとAIQM）も公開された。 MLSEやQA4AIでは実践的な知識やノウハウの共有も進みつつある。しかし、AIの品質や安全性・信頼性を確保し、Trusted Quality AIを日本の強みとして確立し、国際競争力を高めていくためには、国として戦略的取り組みを推進する体制・仕組みをいっそう強化していくことが必要と考える。 AIソフトウェア工学の研究開発は、（1）学術研究と人材育成、（2）実応用での技術実証、（3）基準策定・標準化、という三つの活動を密連携させて推進することが不可欠であり、そのための司令塔の役割を持つ部門が重要になってくる。社会で受容される適切な品質基準・安全性基準を国として策定し、その認証を行う機関を設立・運用（評価のための適切なデータセットの構築・管理も含む）し、標準化活動とも連動させていくことが望まれる。

また、産業界を中心に問題意識が高まり、MLSEを中心に研究コミュニティーも活性化してきたが、その一方で、学術界での取り組みは、まだ一部の研究機関に偏っているように思える。実践に基づく産業界

での取り組みと並行して、パラダイム転換に対する原理・理論の基礎的な研究も強化が望まれる。

❷ 機械学習活用に関わる知的財産権の整備

機械学習に用いるデータや解析結果に関わる知的財産権に加えて、機械学習固有の問題として訓練（学習）済みモデルの知的財産権の問題がある。訓練済みモデルの再利用のパターンは、（1）Copy：そのまま複製して使う、（2）Fine Tuning：ある訓練済みモデルにさらにデータを与えて追加訓練したものを行う、（3）Ensemble：複数の訓練済みモデルの出力を束ねて（平均・多数決など）使う、（4）Distillation：ある訓練済みモデルの振る舞い（どんな入力を与えたときにどんな出力が得られるか）を訓練データとして作ったモデルを使う、という4通りがある[3)]。このようなパターンを含めて、訓練済みモデルの知的財産権をどのように保護すべきか、法整備が必要である。

（7）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	国の「AI戦略2019」や経団連の「AI活用戦略」にTrusted Quality AIが掲げられ、高品質で信頼されるAIを日本の強みとして打ち出そうとする方針が示された。文科省の2020年度戦略目標として「信頼されるAI」が設定され、JSTプログラム（ERATO、MIRAI、CREST、さきがけなど）でAI信頼性に関する研究課題が推進されている。ただし、基礎研究は少数の中核研究者によって牽引されているのが現状で、研究者層がまだ薄い。
	応用研究・開発	○	↗	2018年に機械学習工学研究会MLSEとQA4AIコンソーシアムが発足し、産業界からの多数の参画もあり、活発に活動が進められている。QA4AI・AIQM品質管理のガイドラインが公開された。NEDOプログラムとしてAI信頼性・説明可能AIなどの研究開発が推進されている。
米国	基礎研究	○	↗	DARPAが2017年からXAIプロジェクト、2018年からAssured Autonomyプロジェクトをスタートさせており、基礎研究への比較的大型の政府投資がなされている。
	応用研究・開発	○	↗	Big Tech企業はAI応用システム開発に関する実践的な手法や知見を保有している。米国の第三者認証機関Underwriters Laboratoriesから2020年に自動運転の安全規格UL4600が発表された。
欧州	基礎研究	○	↗	自動運転分野の安全性評価の基準や評価手法の開発のため、ドイツの産官学連携によるペガサスプロジェクトが2016年～2019年に実施された。
	応用研究・開発	○	↗	英国のDeepMindが自社のAI開発ガイドラインをまとめ、公開している。ドイツではDFKIとTÜV SÜDが共同でAIに関する第三者認証の検討を始めた。
中国	基礎研究	○	↗	データ品質やアノテーションに関する品質特性や評価プロセスなど、現実のAIモデルに即した観点からの検討も進められている。
	応用研究・開発	○	↗	多数の中国主要IT企業が参加して、中国のAI国内標準が作られているとともに、国際標準化にも力を入れている。
韓国	基礎研究	×	→	現状、特段の活動が見られない。
	応用研究・開発	×	→	現状、特段の活動が見られない。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

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

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

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

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

参考文献

1）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」, CRDS-FY2018-SP-03（2018年12月）.

2）丸山宏,「機械学習工学に向けて」,『日本ソフトウェア科学会第34回大会講演論文集』(2017年9月）.

3）丸山宏・城戸隆,「機械学習工学へのいざない」,『人工知能』(人工知能学会誌)33巻2号(2018年3月), pp. 124-131.

4）石川冬樹・丸山宏（編著）,『機械学習工学』(講談社, 2022年）.

5）Andrej Karpathy, “Software 2.0”, *Medium*（2017.11.12）. https://medium.com/@karpathy/software-2-0-a64152b37c35（accessed 2023-02-01）

6）Kunle Olukotun, “Designing Computer Systems for Software 2.0”, Invited Talk（December 6, 2018）in *the 32nd Conference on Neural Information Processing Systems*（NeurIPS 2018; Montréal, Canada, December 3-8, 2018）.

7）高野敦,「もうブラックボックスじゃない, 根拠を示してAIの用途拡大」,『日経エレクトロニクス』2018年9月号（2018年）, pp. 53-58.

8）Amina Adadi and Mohammed Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence（XAI）”, *IEEE Access* Vol. 6（17 September 2018）, pp. 52138-52160. doi: 10.1109/ACCESS.2018.2870052

9）Alejandro Barredo Arrieta, et al., “Explainable Artificial Intelligence（XAI）: Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, arXiv: 1910.10045（2018）.

10）Kate Crawford, “The Trouble with Bias”, Invited Talk（December 5, 2017）in *the 31st Conference on Neural Information Processing Systems*（NIPS 2017; Long Beach, Calfornia, December 4-9, 2017）.

11）「AI and bias：人工知能は公平か?」,『MITテクノロジーレビュー Special Issue』Vol. 7（2018年）.

12）神嶌敏弘・小宮山純平,「機械学習・データマイニングにおける公平性」,『人工知能』(人工知能学会誌)34巻2号（2019年3月）, pp. 196-204.

13）Kevin Eykholt, et al., “Robust Physical-World Attacks on Deep Learning Models”, arXiv：1707.08945（2017）.

14）進藤智則,「深層学習や機械学習の品質をどう担保するか?新しいソフト開発手法と位置付け「工学体系」構築へ」,『日経ロボティクス』2018年6月号（2018年）, pp. 3-10.

15）AIプロダクト品質保証コンソーシアム(QA4AIコンソーシアム)編,「AIプロダクト品質保証ガイドライン」(初版2019年5月17日公開, 以降改訂を重ねて最新版2022年7月15日公開）.

16）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（第1版)」, 産業技術総合研究所サイバー

フィジカルセキュリティ研究センターテクニカルレポートCPSEC-TR-2020001（初版2020年6月30日公開, 以降改訂を重ねて最新3.2.1版2023年1月20日公開）.
17）桑島洋・平田雄一・中江俊博,「自動車業界における機械学習システムの品質確保の事例」,『システム/制御/情報』（システム制御情報学会誌）66巻5号（2022年5月）, pp. 187-194. DOI: 10.11509/isciesci.66.5_187
18）本橋洋介,『人工知能システムのプロジェクトがわかる本：企画・開発から運用・保守まで』（翔泳社, 2018年）.
19）Valliappa Lakshmanan, Sara Robinson and Michael Munn, *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLops*（Oreilly & Associates Inc., 2020）.（邦訳：鷲崎弘宜・他3名訳,『機械学習デザインパターン：データ準備、モデル構築、MLOpsの実践上の問題と解決』, オライリージャパン, 2021年）
20）小川雅晴,「AIに関するルール・標準化の動向と今後の展望」, JEITA 国際戦略・標準化セミナー ～Society5.0を創造する新たな標準化の取組み～（2019年10月17日）. https://home.jeita.or.jp/press_file/20191023145047_3Ezs15ATUG.pdf (accessed 2023-02-01)
21）飯泉紀子・鷲崎弘宜・誉田直美（監修）, SQuBOK策定部会（編）,『ソフトウェア品質知識体系ガイド（第3版）－SQuBOK Guide V3－』（オーム社, 2020年）.
22）Jie M. Zhang, et al., “Machine Learning Testing: Survey, Landscapes and Horizons”, *IEEE Transactions on Software Engineering*（Early Access, 17 February 2020）. DOI: 10.1109/TSE.2019.2962027
23）中島震,『ソフトウェア工学から学ぶ 機械学習の品質問題』（丸善出版, 2020年）.
24）Shiqing Ma, et al., “MODE: automated neural network model debugging via state differential analysis and input selection”, *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*（ESEC/FSE 2018, Lake Buena Vista, USA, November 4-9, 2018）, pp. 175-186. DOI: 10.1145/3236024.3236082
25）石本優太・他,「ニューラルネットワークモデルのバグ限局・自動修正技術」,『情報処理』（情報処理学会誌）63巻11号（2022年11月）, pp. e28-e33.
26）Ulrich Aivodji, et al., “Fairwashing: the risk of rationalization”, *Proceedings of the 36th International Conference on Machine Learning*（ICML 2019; June 9-15, 2019）, PMLR 97: pp. 161-170.
27）National Institute of Standards and Technology, “Four Principles of Explainable Artificial Intelligence”, Draft NISTIR 8312（August 2020）. DOI: 10.6028/NIST.IR.8312-draft
28）森川郁也,「機械学習セキュリティ研究のフロンティア」, 電子情報通信学会 基礎・境界ソサイエティ『Fundamentals Review』15巻1号（2021年7月）, pp. 37-46. DOI: 10.1587/essfr.15.1_37
29）日本ソフトウェア科学会機械学習工学研究会,「機械学習システムセキュリティガイドライン」（Version 1.03：2022年12月26日）.
30）SaFAD members（Aptiv, Audi, Baidu, BMW, Continental, Fiat Chrysler Automobiles, Daimler, HERE, Infineon, Intel and Volkswagen）, “Safety First for Automated Driving”（SaFAD White Paper, 2019）.
31）Ichiro Hasuo, et al., “Goal-Aware RSS for Complex Scenarios Via Program Logic”, *IEEE Transactions on Intelligent Vehicles*（July 5, 2022）, pp. 1-33. DOI: 10.1109/TIV.2022.3169762
32）蓮尾一郎,「統計的機械学習と演繹的形式推論：システムの信頼性と説明可能性へのアプローチ」,『日本数学会 2018年度秋季総合分科会 数学連携ワークショップ』（2018年9月24日）. http://group-

mmm.org/~ichiro/talks/20180924okayama.pdf (accessed 2023-02-01)
33) 科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：第４世代AIの研究開発―深層学習と知識・記号推論の融合―」, CRDS-FY2019-SP-08（2020年3月）.
34) 情報処理推進機構 技術本部ソフトウェア高信頼化センター,『はじめてのSTAMP/STPA ～システム思考に基づく新しい安全性解析手法～』『はじめてのSTAMP/STPA（活用編）～システム思考で考えるこれからの安全～』(情報処理推進機構, 2016年3月）.
35) 中江俊博・桑島洋,「自動車業界におけるAIセーフティ動向」,『人工知能』(人工知能学会誌）38巻2号（2023年3月）, pp. 210-220.
36) Hiroshi Kuwajima, Hirotoshi Yasuoka and Toshihiro Nakae, “Engineering problems in machine learning systems”, *Machine Learning* Vol. 109 (April 2020), pp. 1103-1126. DOI：10.1007/s10994-020-05872-w