

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。

(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、ソフトウェアテスト、訓練済みモデル、自動機械学習、AutoML、MLOps、SOTIF、SaFAD

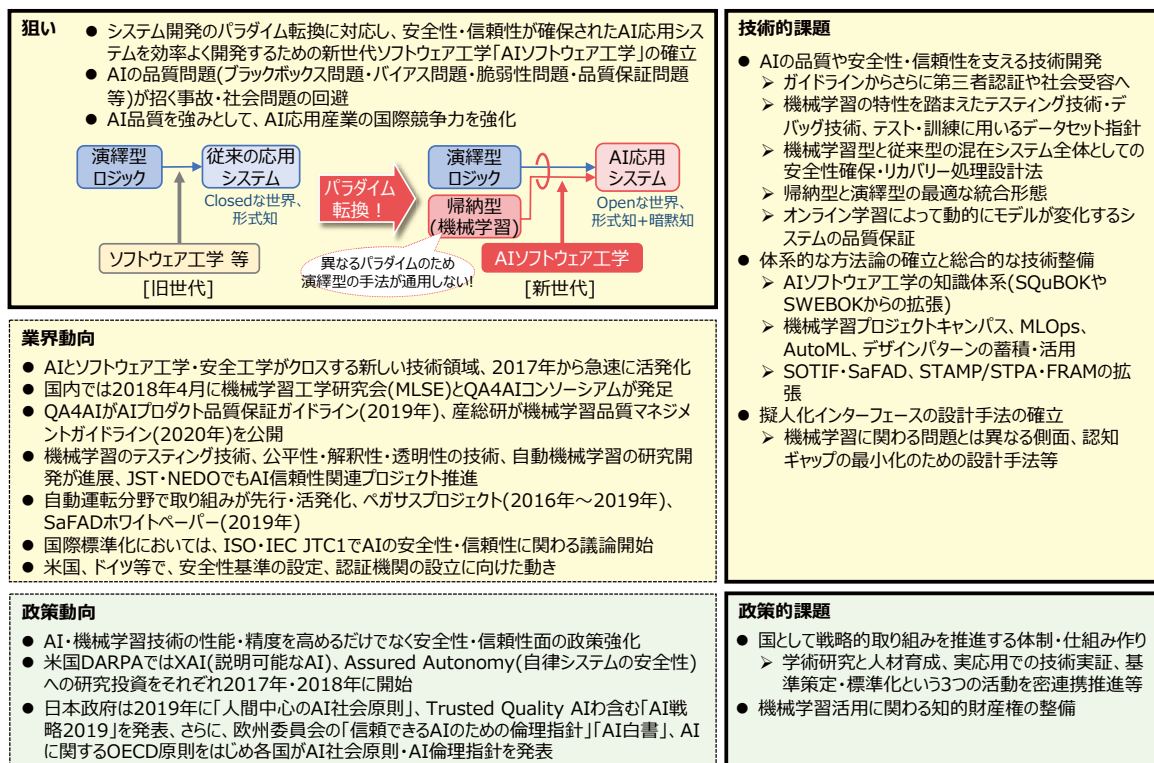


図2-1-6 領域俯瞰：AIソフトウェア工学

(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字モデル等）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テグレーターは競争力を失いかねない。また、動作保証・品質保証等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。事故発生時に原因説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：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等の誤認識を誘発する攻撃²⁾や、悪意をも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判別・分類、予測、異常検知等が可能になった。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、その効率のよい開発を可能にする。

- 1 これはAI応用システムの安全性・信頼性等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・信頼性等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。

[研究開発の動向]

① 学術界・産業界の動向

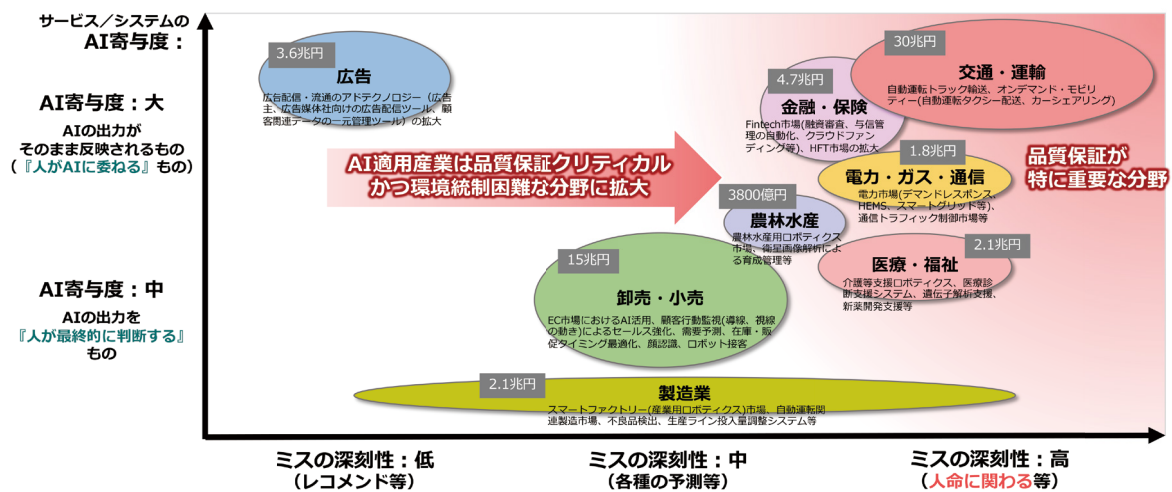
AIソフトウェア工学は、AIと、ソフトウェア工学（Software Engineering）や安全工学（Safety Engineering）がクロスする新しい技術領域である。国内では、2015年頃からシステム開発のパラダイム転換への対応が必要だという問題提起がされ始め、2017年初頭から学会・業界イベント（2017年2月の情報処理学会ソフトウェアジャパン2017、同年8月の情報処理学会ソフトウェア工学シンポジウムSES2017、他多数）で基調講演や企画セッション等が立て続けに開催され、一気にホットトピック化した¹⁾。

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、勉強会等、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⁾。2019年度には、活動のアウトプットを明確化したワーキンググループが5つ立ち上げられた³⁾。

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⁾。産学27の発起人・団体で設立され、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」¹⁵⁾が公開された。また、2020年6月には産業技術総合研究所（AIST）から「機械学習品質マネジメントガイドライン」¹⁶⁾が公開された（これら2つのガイドラインについては〔新展開・技術トピックス〕①を参照）。

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⁾や「Machine Teaching」¹⁷⁾と呼ばれ、国際学会（2018年6月のISCA 2018、2018年12月のNeurIPS 2018等）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¹⁸⁾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。

3 プロセス・事例収集WG、システム基礎WG、本番適用のためのインフラと運用WG、データ品質エンジニアリングWG、機械学習システム セーフティ・セキュリティWGの5つである。

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毎に異なる。応用システムが持つ3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取区分システム等は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。国際的には、ペガサス（Pegasus）プロジェクト¹⁹⁾（2016年1月～2019年6月）、SaFADホワイトペーパー²⁰⁾（Safety First for Automated Driving、2019年6月公開）等、いずれも十数の企業・団体共同で自動運転の安全性が検討されている（詳細は「注目すべき国内外のプロジェクト」①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

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://www.shinnihon.or.jp/shinnihon-library/publications/issue/eyi/knowledge/fsi/pdf/2015-09-15.pdf> (accessed 2021-1-15)

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。

りや技術開発²¹⁾を進めている等、取り組みが活発化している⁶。

また、産業界では、品質管理・安全性確保という面に留まらず、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²²⁾。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」⁷も活用されている。技術的には、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（その詳細は「新展開・技術トピックス」④を参照）。

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、ここ数年で社会における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等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等）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（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习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。

② 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に関わる動きが生まれている。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²³⁾。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）の第一合同技術委員会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始まっている。特に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のガバナンス上の意義に関する検討を行うSC40との合同WGが活動しており（WG：Working Group）注目される。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

6 国内における自動運転に関わる制度整備の状況や官民連携プロジェクト（SIP-sdus）については「2.1.9 社会におけるAI」の中に記載している。

7 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パス」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。

もIECでは標準管理評議会（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等を目的に設置された。

自動運転の安全規格については、米国の第三者認証機関であるUnderwriters Laboratoriesが2020年4月にUL4600を発表した。これは完全に自律的に走行する車両の安全性を評価するための原則とプロセスを示しているが、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。

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²⁴⁾。

③ 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」は、人間の尊厳が尊重される社会（Dignity）、多様な背景を持つ人々が多様な幸せを追求できる社会（Diversity & Inclusion）、持続性ある社会（Sustainability）を基本理念に定め、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⁸として7つのAI社会原則を掲げた。7つは、（1）人間中心の原則、（2）教育・リテラシーの原則、（3）プライバシー確保の原則、（4）セキュリティ確保の原則、（5）公正競争確保の原則、（6）公平性、説明責任および透明性の原則、（7）イノベーションの原則である。「AI戦略2019」ではAI社会原則を受けて、「Trusted Quality AI」⁹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推進され

8 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」等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、プライバシー、倫理、利用者支援、アカウントビリティという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活用ガイドライン」も公表した。

9 「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。

ている（〔注目すべき国内外のプロジェクト〕③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の「人工知能に関するOECD原則」(OECD Principles on Artificial Intelligence、2019年5月、42か国が署名)、中国の「次世代AIガバナンス原則—責任あるAIの発展」(新一代人工智能治理原則—発展負責任的人工智能、2019年6月)等が挙げられる（詳細および関連動向は「2.1.9 社会におけるAI」参照）。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）プロジェクトとAssured AutonomyプロジェクトがAIの安全性・信頼性に関するものとして知られている（これらの詳細は〔注目すべき国内外のプロジェクト〕②に記載）。

（4）注目動向

〔新展開・技術トピックス〕

① AI 品質管理ガイドライン

〔研究開発の動向〕①で述べたように、AI 品質管理に関する問題意識を背景に、2018年からAI 品質管理ガイドラインを作ろうという活動が立ち上がった。国内では、特にQA4AI コンソーシアムによる「AI プロダクト品質保証ガイドライン」(2019年5月に初版公開、2020年2月・8月にアップデート)とAISTの「機械学習品質マネジメントガイドライン」(2020年6月に初版公開)がよく知られている。

QA4AIのAI プロダクト品質保証ガイドライン¹⁵⁾では、品質保証で考慮すべき軸としてData Integrity、Model Robustness、System Quality、Process Agility、Customer Expectationという5つが定義され、それぞれに関してチェックリストが示されている。そして、生成系システム、スマートスピーカー（Voice User Interface）、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。また、品質保証に関わる技術カタログや、機械学習における説明可能性・解釈性に関する技術動向も示されている。

AISTの機械学習品質マネジメントガイドライン¹⁶⁾では、機械学習を利用したAIシステムにおける品質を、AIシステム利用時に必要な利用時品質、AIシステム中で機械学習要素に要求される外部品質、機械学習要素そのものが持つ内部品質の3つに分け、機械学習要素の内部品質の向上により外部品質を必要なレベルで達成し、最終的な製品の利用時品質を実現するという考え方が示されている。さらに、機械学習要素に特有の外部品質として、リスク回避性、AIパフォーマンス、公平性があるとし、AIシステムの利用時品質もこの3つを基本として表現されるとした。一方、内部品質は、要求分析の十分性、データ設計の十分性、データセットの被覆性、データセットの均一性、機械学習モデルの正確性、機械学習モデルの安定性、プログラムの健全性、運用時品質の維持性という8項目に整理されている。これらの内部品質をプロセス管理や数値評価によって具体的に確認して、外部品質に求められる要件を達成することを目指す、その過程での考え方や参照モデルも示されている。

QA4AIのガイドラインは開発者にとってのチェックリスト的な位置付けで活用でき、一方、AISTのガイドラインは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、2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品質管理の意識は高まっている。また、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、分野毎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¹⁰。

② 機械学習のテスト技術

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」等、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（ごく小さなノイズ等）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。

機械学習のテスト技術は、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²⁵⁾。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^{25), 26)}。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ト、サーチベースドテスト、データセット多様性等のアイデアが知られている。ニューロンカバレッジは、深層学習等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トは、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トは、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。このようなアイデアの意義・有効性や限界は、まだ必ずしも見極められてはおらず、今後の研究進展や開発現場適用を通して示されていくものである。

さらに、機械学習単体でのテストではなくシステム全体での評価が検討されるようになった。機械学習による認識ミスが必ずしもシステムの事故や要求違反になるとは限らないため、状況の起きやすさや問題の大きさに基づくリスク評価を考える必要がある。エッジケースを想定したSOTIF (Safety Of The Intended Functionality、ISO DIS 21448) の考え方等が取り入れられてきている。また、問題を見つけるだけで

10 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 2021-01-15)

なく原因把握・デバッグの手法も検討されるようになった²⁷⁾。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等が提案されている。

③ 機械学習における公平性・解釈性・透明性 (FAT/ML)

上記②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、規制遵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、FADM (Fairness-aware Data Mining : 公平性配慮データマイニング)¹²⁾ と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等によって、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・個人公平性等の公平性規準を定義し、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。ただし、公平性を確保することで通常、精度は低下するので、応用毎の要件を明確化して設計することが必要である。

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1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等)。もう1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等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等、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・信頼の確保等、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、騙すような説明も作れてしまう可能性がある。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

2.1

俯瞰区分と研究開発領域 人工知能・ビッグデータ

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²⁸⁾。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。

④ 自動機械学習 (AutoML)

OSS化も進み、機械学習は広く使える道具になったが、この道具を使いこなせる人材 (いわゆるデータサイエンティスト) は依然として不足している。機械学習はその構造設計やパラメーター設定等にノウハウや試行錯誤が必要となることが多く、道具があっても使いこなすのは必ずしも容易ではない。また、大きな計算リソースを必要とする機械学習をクラウドで利用したいというニーズが高まっている。そこで、データと目的を与えれば結果が自動で出てくるような、機械学習の設計自動化、自動機械学習の道具・フレームワークを目指した取り組みが活性化し、商用提供も始まった。これによって、データサイエンティスト不足が補われ、産業応用が広がることが期待される。

2012年に米国ボストンで創業したDataRobotは、データサイエンティストの知識・ノウハウをソフトウェア化し、深層学習、決定木、ナイーブベイズ、SVM等、100以上のアルゴリズムから、問題やデータの性質に合う複数アルゴリズムを試し、精度の高い順にモデルを提示する機能を提供している。NECは2016年にリレーショナルデータベースでの大規模データ予測分析プロセス全体を自動化する予測分析自動化技術 (特徴量自動設計技術+予測モデル自動設計技術) を発表し、2017年にこの技術を製品提供する新会社dotDataを北米に設立した。銀行関連のデータ分析業務への適用事例では、従来2~3か月を要した作業期間が1日に短縮できたとのことである。

さらにGoogleはクラウド上で、Cloud Vision APIやNatural Language API等、訓練 (学習) 済みモデルによるサービスを提供してきたことに加えて、ユーザーが用意した訓練 (学習) データを用いる自動機械学習サービスCloud AutoMLを発表した。2018年1月に画像解析 (Cloud AutoML Vision) がまずリリースされ、これまでに自然言語処理 (Cloud AutoML Natural Language)、機械翻訳 (Cloud AutoML Translation)、動画解析 (Cloud AutoML Video Intelligence)、構造化データ解析 (Cloud AutoML Tables) もリリースされている。これらは、訓練済みモデルを別ドメインに転用して再学習する転移学習 (Transfer Learning) 技術と、対象ニューラルネットワークの構造やハイパーパラメーターを別のニューラルネットワークを用いた強化学習によって最適化探索するNAS (Neural Architecture Search) 技術によって実現されている²⁹⁾。他にもMicrosoftのAzure Machine Learning、H2O Driverless AI他、各社からAutoMLサービスが提供されている。

[注目すべき国内外のプロジェクト]

① 自動運転の AI 安全性に関するプロジェクト

ドイツ経済エネルギー省 (Bundesministerium für Wirtschaft und Energie : BMWi) が主導するベガサスプロジェクト¹⁹⁾ は、自動走行システムの期待性能水準 (自動運転車両はどれくらいの性能が求められるか?) と評価基準 (求められる性能を満たすことをどのようにして確認するか?) を明確化するための産学官共同プロジェクトとして、2016年1月~2019年6月に実施された。BMW、Daimler、Volkswagen、Audi等の自動車メーカー、Robert Bosch GmbH (ボッシュ) 等のサプライヤー、ドイツ

航空宇宙センターやダルムシュタット工科大学等の大学・研究機関、第三者認証機関TÜV SÜD（テュフズード）等、計17団体が参加した。従来のV字モデルに代わる自動運転の安全性評価フレームワークについて、自動走行車の機能・要件定義、走行環境・走行状況のデータ化、それらを組み合わせたテストシナリオ作成、テストシナリオに基づくシミュレーション・実走行テストとリスク許容性を踏まえた安全性評価という4項目が検討された。ただし、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、自動運転システムの中で使われているAI技術そのものの安全性要件を明に論じてはいない。

また、2019年6月にSaFADホワイトペーパー²⁰⁾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ISO TR4804が発行された。

一方、欧州委員会は2020年2月に「AI白書」(White Paper on Artificial Intelligence - A European approach to excellence and trust)を発表し、市民の価値観と権利を尊重した安全なAI開発の「信頼性」と「優越性」を実現するための政策オプションを示した。これに対して欧州自動車工業会(ACEA)は、11月に「Position Paper: Artificial Intelligence in the automobile industry」を発表した。欧州委員会が水平的(横断的)アプローチでAI要件の立法化を検討しているのに対して、ACEAは産業セクターで定められた既存の枠組み(型式認証等)に新たなAI要件を盛り込むべきとするセクター別アプローチを主張している。

自動運転システムの安全開発に関する主要な国際規格にはISO 26262とSOTIF (ISO DIS 21448)がある。ISO 26262はハードウェア故障や仕様逸脱に対する安全性(機能安全)を対象とするが、それだけでは不十分で、SOTIF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する。これは、未知のシナリオと危害が及ぶシナリオを特定し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。このSOTIFの考え方は、自動運転におけるAI安全性確保の主流になりつつある。

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等)。自動車業界と学術界の両面で活発に取り組まれている。

② 米国 DARPA における関連プロジェクト (XAI、Assured Autonomy)

米国DARPAのAI関連プログラムの中で、安全性・信頼性に関わり、特によく知られているのがXAIプロジェクトである。比較的早い時期にスタートしており(プロジェクト期間は2017年5月~2021年4月の4年間)、XAIという名称はこのプロジェクトから使われるようになったものである。このプロジェクトは、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ている。つまり、技術的には、機械学習の精度と解釈性の両立が目標であり、具体的なタスクとしてData AnalyticsとAutonomyという2つを設定している。Data Analyticsタスクは、マルチメディアデータを解析してターゲットを選択するに際して、選択の理由も示す(例えば、敵と判断した理由として銃の輪郭を強調表示する等)。Autonomyタスクは、ドローンやロボット等の自律システムにおいて、どういう状況でど

という理由で次の行動を決定したかを説明する。XAI技術の種類については「新展開・技術トピックス」③で述べたが、このプロジェクトでも複数通りのアプローチで取り組まれている。また、説明可能なモデルだけでなく、説明のためのインターフェースや説明の心理学的モデルの研究にも取り組まれている。

もう一つ、Assured Autonomyプロジェクト（2018年5月～2022年4月の4年間）では、自動運転車やドローン等の自律システムの安全性確保が検討されている。そのために、実行時の入出力をモニタリングし、想定された振る舞いから外れたときに、それを検知して動作を止める等の対策アクション（事前に設計しておいたもの）を起動するような仕組みが研究開発されている。

③ 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」③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年度は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」¹¹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～2022年3月、研究総括：蓮尾一郎）が、高信頼自動運転システムを重点ターゲットとして、物理情報システムの品質保証手法の学術的研究に取り組んでいる。また、未来社会創造事業「超スマート社会の実現」領域に「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）が採択され、現在は「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）に移行している。さらに、文部科学省の2020年度戦略目標の1つとして「信頼されるAI」が定められ、それを受けた「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）、「さががけ 信頼されるAIの基盤技術」（研究総括：有村博紀）がスタートした。

（5）科学技術的課題

① 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。QA4AIやAISTによって品質管理ガイドラインがまとめられたが、チェックリストやユースケースが中心であり、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、というところさらに実績を積みながら内容を整備していくことが必要である。ガイドラインで示された品質評価軸に沿った機械学習の特性を踏まえたテスト技術やデバッグ技術の強化、そして、テストや訓練に用いるデータセットの質・量についての考え方・指針等も求められる。

11 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。

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 Envelope³⁰⁾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³¹⁾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。

② 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、前述したように機械学習プロジェクトキャンパスの活用やMLOpsも実践されている。開発効率化の面では、[新展開・技術トピックス] ④に示した自動機械学習（AutoML）技術のほかにも、設計ノウハウをカタログ化したデザインパターンを蓄積・活用することも考えられている³²⁾。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^{33), 34)} (System Theoretic Accident Model and Processes / System Theoretic Process. Analysis) やFRAM (Functional Resonance Analysis Method) 等のAI・機械学習の応用システムへの拡張適用¹⁴⁾等も検討されている。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。

③ 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(詳細は「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。

(6) その他の課題

① 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〕③や〔注目すべき国内外のプロジェクト〕③で述べたように、国のAI 戦略の中でも言及され、NEDOやJSTのプログラムも立ち上がった。QA4AIやAISTから産業界で活用できるガイドラインも公開された。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3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も活性化してきたが、その一方で、学術界での取り組みは、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変更して、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。

② 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・多数決等) 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、さががけ等）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とAISTからAI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等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	○	↗	GAFA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	○	↗	イギリスの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	△	→	AI分野の研究論文数は米中が圧倒的で、中国政府も「次世代AIガバナンス原則」を発表したが、AIの安全性・信頼性に関わる基礎研究への積極的投資は見られない。
	応用研究・開発	△	→	欧米企業に比べると、安全性・信頼性に関する中国企業での取り組み意識は弱い。
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) 野々山尊秀・丸山宏 (編), 「特集: 機械学習工学」, 『情報処理』(情報処理学会誌) 60巻1号 (2019年1月), pp. 10-55.
- 5) Andrej Karpathy, “Software 2.0”, *Medium* (2017.11.12) . <https://medium.com/@karpathy/software-2-0-a64152b37c35> (accessed 2020-11-15)
- 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, California, 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) 神島敏弘・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」(2020年8月版) .
- 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 (2020年6月30日) .
- 17) Patrice Y. Simard, et al., “Machine Teaching : A New Paradigm for Building Machine Learning Systems”, arXiv : 1707.06742 (2017) .
- 18) Pedro A. Ortega, Vishal Maini and the DeepMind safety team, “Building safe artificial intelligence : specification, robustness, and assurance”, *Medium* (2018.9.28.)
<https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1> (accessed 2020-11-15)
- 19) Ulrich Eberle, et al., “Automated Driving & Development Processes-A Brief Look Beyond”,

Pegasus Symposium 2017 (Aachen, Germany, November 9, 2017) .

- 20) 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) .
- 21) 桑島洋・安岡宏俊・中江俊博, 「セーフティクリティカルな機械学習システムの品質保証の課題」, 『第1回機械学習工学ワークショップ論文集』(MLSE2018;2018年7月) .
- 22) 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) .
- 23) 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 2021-1-15)
- 24) 飯泉紀子・鷺崎弘宜・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) .
- 25) 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
- 26) 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) .
- 27) 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
- 28) National Institute of Standards and Technology, “Four Principles of Explainable Artificial Intelligence”, Draft NISTIR 8312 (August 2020) . DOI : 10.6028/NIST.IR.8312-draft
- 29) 森正和, 「最新AI技術「NAS」(Neural Architecture Search) を紐解く ～Google Cloud AutoMLの裏側を見ていく～」, 『Developers Summit FUKUOKA 2018』(福岡,2018年9月6日) .
- 30) 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 2020-11-15)
- 31) 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の研究開発ー深層学習と知識・記号推論の融合ー」, CRDS-FY2019-SP-08 (2020年3月) .
- 32) 鷺崎弘宜, 「機械学習システムの33のアーキテクチャパターンおよびデザインパターン」, スマートエスイーセミナー: 機械学習デザインパターンとアジャイル品質パターン (2019年11月8日) . <https://www.slideshare.net/hironoriwashizaki/ss-191808794> (accessed 2021-1-15)
- 33) 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～』(情報処理推進機構,2016年3月) . <https://www.ipa.go.jp/files/000051829.pdf> (accessed 2020-11-15)
- 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 (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) . <https://www.ipa.go.jp/files/000065199.pdf> (accessed 2020-11-15)

2.1

俯瞰区分と研究開発領域
人工知能・ビッグデータ