CRDS-FY2018-SP-03

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

STRATEGIC PROPOSAL

Toward new-generation software engineering to guarantee safety and reliability of AI application systems

国立研究開発法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査、分析、提案を中立的な立場に立って行う公的シンクタンクの一つで、文部科学省を主務省とする国立研究開発法人科学技術振興機構（JST）に属しています。

CRDS は、科学技術分野全体像の把握（俯瞰）、社会的期待の分析、国内外の動向調査や国際比較を踏まえて、さまざまな分野の専門家や政策立案者との対話を通じて、「戦略プロポーザル」を作成します。

「戦略プロポーザル」は、今後国として重点的に取り組むべき研究開発の戦略や、科学技術イノベーション政策上の重要課題についての提案をまとめたものとして、政策立案者や関連研究者へ配布し、広く公表します。

公的な科学技術研究は、個々の研究領域の振興だけでなく、それらの統合によって社会的な期待に応えることが重要です。「戦略プロポーザル」が国の政策立案に活用され、科学技術イノベーションの実現や社会的な課題の解決に寄与することを期待しています。

さらに詳細は、下記ウェブサイトをご覧下さい。

http://www.jst.go.jp/crds/about/

エグゼクティブサマリー

現在、人工知能（AI：Artificial Intelligence）の第 3 次ブームと言われるが、これを牽引しているのは、深層学習（ディープラーニング：Deep Learning）をはじめとする機械学習技術である。そして、研究コミュニティーや新聞・雑誌メディアを賑わせているだけでなく、機械学習技術は様々なシステムに埋め込まれ、実社会での応用・実用化が急速に広がっている。

この機械学習技術はシステムの動作を大量データから帰納的に定義するものであり、プログラムコードを書いてシステムの動作を定義する従来の開発スタイルとは大きく異なっている。そのため、従来ソフトウェア工学として構築・整備されてきた方法論・技術は必ずしも使えず、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。

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の社会実装が急激に進むと、そこで発生した問題や事故が思わぬ社会問題化する懸念もある。

このような産業界にも大きなインパクトを与え得るシステム開発方法のパラダイム転換に際し、AI･機械学習技術そのものの研究開発が強力に推進されているのに比べて、それを組み込んだ応用システムの開発のための技術・方法論の研究開発については、まだ取り組みが十分でない。

そこで、本プロポーザルでは「AI ソフトウェア工学」の確立を目指し、その研究開発を推進強化することを提案する。

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するための技術･方法論がソフトウェア工学の中で整備されてきた。それに対し、AI ソフトウェア工学は、いわば新世代のソフトウェア工学であり、従来型だけでなく、機械学習型のコンポーネントも含む AI 応用システムを対象として、その安全性・信頼性を確保するための技術・方法論を意味する。

なお、本プロポーザルでは、この新世代ソフトウェア工学を「AI ソフトウェア工学」と呼ぶが、国内では「機械学習工学」、海外では「Software 2.0」等と呼ばれるものが、これにほぼ相当する。

AI ソフトウェア工学を新しい学問分野として発展･確立させ、わが国がこの技術で国際競争力を確保するための研究開発課題を、2 つの面からあげる。1 つは「AI ソフトウェア工学の体系化」であり、もう 1 つは「基礎研究として重要な技術チャレンジ」である。

まず、システムの安全性・信頼性の確保は、1 つの技術によって達成し得るものではなく、多面的・総合的な取り組みが不可欠である。全体像を押さえつつ、見通しよく、バランスよく、取り組むべきである。そのため、個々の技術開発と並行して、全体像としての技術体系を描きつつ進めるという意味で、「AI ソフトウェア工学の体系化」を研究課題と

してあげた。

また、AI ソフトウェア工学に対する多面的・総合的な取り組みは、（A）実践から得られる知見を体系化していくアプローチや（B）理論・原理に関わる学術研究のアプローチが可能である。これらのうちアプローチ（A）は産業界を中心に自ずと進んでいくに違いない。その一方で、システム開発のパラダイム転換に対応するには、より根本的な取り組みとして、アプローチ（B）が不可欠である。

産業界を中心としたアプローチ（A）に委ねるだけでなく、国としては、学術研究のアプローチ（B）についても重点推進すべきというのが､2 つめの研究開発課題「基礎研究として重要な技術チャレンジ」をあげた理由である。そして、特に重要な技術チャレンジとして、①機械学習自体の品質保証、②全体システムとしての安全性確保、③ブラックボックス問題への対策、④問題を効率よく解く工学的な枠組みという 4 つをあげる。

また、基礎研究課題の他に、国として検討が求められる課題として、安全性基準の策定（および第三者評価の仕組み）、訓練済みモデルの再利用に関わる知財権等の検討も並行して進めることが求められる。

このような AI ソフトウェア工学の研究開発を推進するためには、新たな施策を講じていくことが必要になる。効果的な施策の設計・実行のためには、以下のような方針・考え方が鍵となる。

- 問題解決型の基礎研究をトップダウンに加速
- 人工知能×ソフトウェア工学の技術融合を促進し、新しい技術分野を創出、事業競争力を強化
- ①学術研究＆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
- 競争が始まっている自動運転分野で先行実証し、他分野へ展開
- 日本の先行優位性を活かす迅速な施策推進

以上、社会の様々なシステムに AI･機械学習技術が組み込まれていく中で、それらシステムの安全性・信頼性を確保し、また、AI・機械学習の品質を強みとして日本の国際競争力を高めていくために、AI ソフトウェア工学の推進・確立が不可欠である。

Executive Summary

We are now in the third Artificial Intelligence（AI）boom. What is driving this boom is "Deep Learning" and other forms of machine learning technology. Machine learning technology does not only excite the research community and print media. It is also embedded into various systems and rapidly put into practical use in the real world.

Machine learning technology inductively defines a system's operation from vast amounts of data, and thus it differs greatly from the conventional development style whereby a program code is written and then the system is defined accordingly. Therefore, the methodologies and technologies that have been built and developed for conventional software engineering are not always applicable. This makes a new way of thinking necessary for the definition of system requirements, performance guarantee, and quality assurance. Hence a paradigm shift is occurring in the system development.

With this paradigm shift, the skills and methodologies required for system development need to be renewed, and the software industry and system integrators may lose its competitiveness if they fail to keep up with this shift. Moreover, if systems embedded with machine learning technology rapidly appear in society, even as approaches to performance guarantee and quality assurance remain undeveloped, the resulting difficulties and accidents could produce unexpected social problems.

The paradigm shift in system development that is described here could have major implications for the industrial world. However, in comparison to R&D on AI and machine learning technology themselves, R&D to methodologies for developing the systems remains insufficient.

In light of this, we, through this proposal, aim to establish "AI software engineering" and recommend the promotion and strengthening of R&D for this purpose.

In the case of conventional system development, technologies and methodologies for maintaining safety and reliability and developing systems efficiently have been established within software engineering. On the other hand, AI software engineering could be described as "next-generation software engineering." It targets not only conventional systems but also systems that apply AI, including machine learning-type components, and refers to technologies and methodologies for maintaining safety and reliability in them.

It should be noted that, while we refer to this kind of next-generation software engineering as "AI software engineering" in this proposal, it roughly corresponds to what is known as "machine learning systems engineering" in Japan and as "Software 2.0" in other countries.

We view R&D themes for developing and establishing AI software engineering as a new academic field and for Japan's attainment of international competitiveness

in this technology from two aspects. The first is the "systematization of AI software engineering" and the second is "important technical challenges as basic research."

First, system safety and reliability cannot be achieved with just one technology; they must be attained through multifaceted and comprehensive initiatives. They must be tackled with clear vision and good balance and with a grasp of the whole picture. Thus, we put forth "the systematization of AI software engineering" as a research theme in the sense of preparing a technical system with a panoramic view and in parallel with the development of individual technologies.

Additionally, multifaceted and comprehensive initiatives to AI software engineering could make the following approaches: involving the systematization of knowledge gained from practical application ("Approach A"), and an approach involving academic study concerning theory and principles ("Approach B"). Of these, Approach A will certainly take place naturally, particularly in industry. On the other hand, Approach B is essential as a more fundamental way of dealing with the paradigm shift in system development.

Our belief that, rather than leaving things to Approach A, which is primarily focused on industry, the national government should also put priority on promoting Approach B for academic research. This the reason why we put forth "important technical challenges as basic research" as the second R&D theme. Here, we present the following four challenges as being particularly important: (1) quality assurance in machine learning in and of itself; (2) ensuring safety in the entire system; (3) tackling the "black box" problem; and (4) design of an engineering framework for efficiently resolving problems.

Furthermore, in addition to basic research topics, the government should simultaneously study the formulation of safety standards (and mechanisms for third-party evaluation) and intellectual property rights concerning the reuse of trained models.

New management measures must be implemented to promote this kind of R&D in AI software engineering. The following policies and approaches will be the key to designing and executing effective measures.

- Top-down acceleration of problem-solving-oriented basic research
- Promotion of the technical fusing of AI and software engineering, creation of new technical fields, and reinforcement of business competitiveness
- Promotion based on the close linkage of three activities; namely, (1) academic research and human resources development, (2) technical validation in practical application, and (3) standardization
- Advance validation in the emerging competitive field of automated driving and expansion to other fields
- Prompt promotion of measures that take advantage of Japan's existing superiority

As was stated above, at a time when AI and machine learning technologies are becoming embedded in the various systems of society, the promotion and establishment of AI software engineering are essential for securing the safety and reliability of those systems and for raising Japan's international competitiveness by using the quality of its AI and machine learning as an advantage.

目　　次

エグゼクティブサマリー
Executive Summary

１．研究開発の内容 …… 1
２．研究開発を実施する意義 …… 5
　２－１．現状認識および問題点 …… 5
　２－２．社会・経済的効果 …… 12
　２－３．科学技術上の効果 …… 12
３．具体的な研究開発課題 …… 14
４．研究開発の推進方法および時間軸 …… 23
付録１．検討の経緯 …… 28
付録２．国内外の状況 …… 35
付録３．専門用語説明 …… 38
付録４．参考文献 …… 40

【コラム】
①「AI ソフトウェア工学」という名称についての補足
②機械学習の簡単なイントロダクション
③ DARPA の XAI プロジェクト
④もう一つの課題：擬人化インタフェース設計

1．研究開発の内容

本プロポーザルで提言をまとめた「AI ソフトウェア工学」は、AI 応用システムの安全性･信頼性を確保する新世代のソフトウェア工学を指す。すなわち、プログラム（手続き）を書くという従来の（演繹型の）システム開発方法に対して、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）システム開発方法についても、安全性・信頼性を確保できる方法論・技術体系である。従来の演繹型システム開発のためのソフトウェア工学に対して、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新世代のソフトウェア工学を、本プロポーザルでは「AI ソフトウェア工学」と呼ぶ（図 1）。

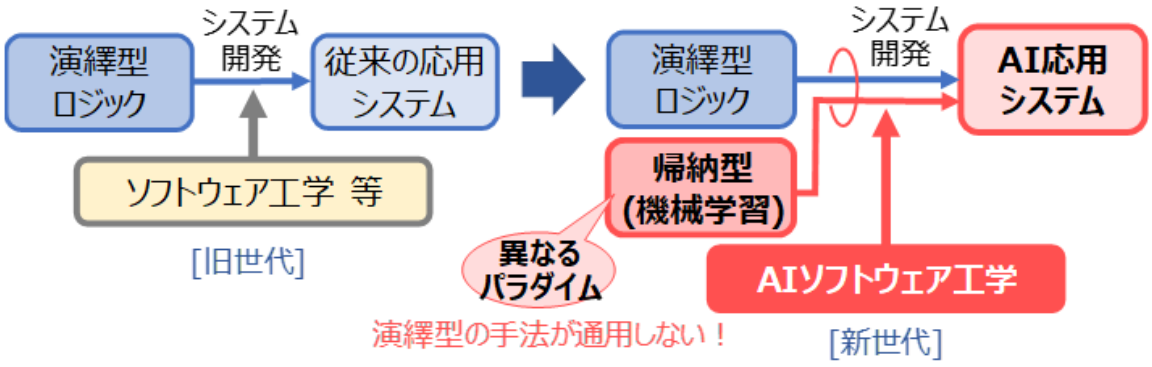


図 1　新世代ソフトウェア工学

AI ソフトウェア工学を新しい学問分野として発展・確立させ、わが国がこの技術で国際競争力を確保するための研究開発課題は、2 つの面からあげられる。1 つは「AI ソフトウェア工学の体系化」であり、もう 1 つは「基礎研究として重要な技術チャレンジ」である。

まず、システムの安全性・信頼性の確保は、1 つの技術によって達成し得るものではなく、多面的・総合的な取り組みが不可欠である。全体像を押さえつつ、見通しよく、バランスよく、取り組むべきである。そのため、個々の技術開発と並行して「AI ソフトウェア工学の体系化」を進めることが重要である。

また、AI ソフトウェア工学に対する多面的・総合的な取り組みは、（A）実践から得られる知見を体系化していくアプローチや（B）理論・原理に関わる学術研究のアプローチが可能である。これらのうちアプローチ（A）は産業界を中心に自ずと進んでいくに違いない。その一方で、システム開発のパラダイム転換に対応するには、より根本的な取り組みとして、アプローチ（B）が不可欠である。

本プロポーザルでは、アプローチ（B）の学術的な基礎研究として重要な技術チャレンジの検討も行い、①機械学習自体の品質保証、②全体システムとしての安全性確保、③ブラックボックス問題への対策、④問題を効率よく解く工学的な枠組みという 4 つの重要な技術チャレンジを抽出した。また、基礎研究課題の他に、国として検討が求められる課題として、安全性基準の策定（および第三者評価の仕組み）、訓練済みモデルの再利用に関わる知財権等の検討も並行して進めることが求められる。

以下、4 つの重要な技術チャレンジの概要を述べる。

①機械学習自体の品質保証

データ集合に着目した機械学習の品質（精度、公平性等）の理論やテスト手法が必要である。機械学習の精度や想定外リスクを左右するのは訓練データ集合なので、データ集合の良さに着目して評価するアプローチが検討されている。また、結果の公平性や訓練済みモデルの再利用性も、データ集合の良さから考えることが有効と考えられる。

②全体システムとしての安全性確保

従来型と機械学習型の混在システムの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。機械学習型コンポーネントが 100% 保証はできないことを想定した全体システム設計や、機械学習型の入力・出力をモニタリングして、例外処理やリカバリー処理・多重化等を組み込むことが有効である。

③ブラックボックス問題への対策

機械学習応用に不安を与えている要因として、機械学習が判定したとき、その理由を人間が理解できないというブラックボックス問題がある。(すべての機械学習方式がブラックボックスというわけではないが）システムにブラックボックスな部分があると、設計時や利用時に、設計者あるいは利用者がシステムの動作を予測することが難しく、システムの安全性を確保する上で妨げになる。また、もしシステムが事故を起こしてしまった場合、その原因理解や責任の切り分けが難しいということにもなる。そこで、システム動作の解釈性を向上させ、事故発生時の原因理解や設計時・利用時の動作予測の容易にすることが求められる。

④問題を効率よく解く工学的な枠組み

機械学習（ボトムアップ）と記号推論（トップダウン）を統合した効率的な問題解法を追求する。深層学習が広がったことで帰納型・ボトムアップ型の問題解決が主流になっているが、本来、演繹型・トップダウン型の記号推論も重要である。それらのあるべき組み合わせや新しい問題解決の枠組みを探究することは、AI 応用システムのアーキテクチャー設計という面で重要である。

さらに、このような基礎研究を中心とした学術研究と人材育成を進めるとともに、産業界と共同で実応用での技術実証、および、AI の品質･安全性に関する基準策定と標準化も並行して推進していくことが、AI 適用産業において AI 品質を強みとした国際競争力の確保につながる。

【コラム①】「AI ソフトウェア工学」という名称についての補足

本プロポーザルでは、新世代のソフトウェア工学に対して「AI ソフトウェア工学」という名称を用いた。これは新しい学問分野・技術体系に対する名称であり、既に業界・研究コミュニティーにおいては「機械学習工学[1][2]」「Software 2.0[3]」「Machine Teaching[4]」等と呼ばれている概念とほぼ一致する。国内外の動向は２－１－４節や付録２に載せているが、国内では「機械学習工学」という名称が定着しつつあり、海外では「Software 2.0」という名称が用いられることが多い。「機械学習工学」は Preferred Networks の丸山宏による命名で、2018 年 4 月には日本ソフトウェア科学会に機械学習工学研究会（Machine Learning Systems Engineering、略称：MLSE、メルシー）が発足し、活発に活動している。「Software 2.0」は Tesla の AI 部門長である Andrej Karpathy が概念を提唱し、最近 1 年間は国際学会の基調講演・招待講演等でこの概念が取り上げられることも多く、認知が広がっている。しかし、新しい概念であるため、その定義やスコープは必ずしも確立されていない。このような状況で本プロポーザルに「機械学習工学」や「Software 2.0」といった名称を冠した場合、それらに独自の解釈を与えることになってしまい、混乱を招くことが危惧された。そこで、本プロポーザルでは、あえて「AI ソフトウェア工学」という新たな名称を用いた。

ただし、「AI ソフトウェア工学」という名称についても 2 つの点で議論がある。

1 点目は「AI」という言葉は世の中で多様な解釈があり、定義やスコープが曖昧になるという懸念である。実際、本プロポーザルでは機械学習にフォーカスして論じているので、技術的・工学的な関心からは「AI」ではなく「機械学習」を言葉として用いた方が正確である。ただ、国の施策を論じるという観点から「AI」を冠した方が位置付けを示しやすい。また、機械学習以外にも、AI 応用システム開発で新たな問題を生む要因があり得ることから、スコープを広めにとっておくことは意味がある。（その新たな問題を生む要因に関しては、コラム④で取り上げている）

2 点目は「ソフトウェア工学」ではスコープが狭いという懸念である。AI 応用システムの安全性・信頼性を確保するためには、ソフトウェア系（従来型および機械学習型）だけでなく、ハードウェア系も含めたシステム全体を対象とし、開発プロセス、機能安全、品質保証を総合的に扱う必要がある。したがって、技術分野としては、狭い意味でのソフトウェア工学にとどまらず、広く安全工学やシステム工学もスコープに含めて取り組むのがよい。そこで、本プロポーザルではソフトウェア工学をやや広めにとらえて体系化していこうという立場である。「ソフトウェア工学」という言葉を別の言葉に置き換えることも検討したが、海外で「Software 2.0」と呼ばれているように、やはり中心はソフトウェアにある。

【コラム②】機械学習の簡単なイントロダクション

機械学習は、データの背後に潜む規則性を発見することにより、人間と同程度あるいはそれ以上の学習能力をコンピューターで実現しようとする技術である。これを用いると、過去の事例・観測データから発見した規則性に基づく、モノやコトの判別・分類、予測、異常検知等の知的な判断をコンピューターで実現可能になる[5]。

機械学習は大きく2つの処理ステップに分かれる。図2に示すように、(A) 訓練ステップと (B) 判定ステップである。(A) は、訓練用の大量データから規則性を発見し、その規則性を表現したモデル（訓練済みモデル）を出力する。(B) は、新たな入力データを、訓練済みモデルと照合することによって、判別・分類、予測、異常検知等の判定を行い、判定結果を出力する。

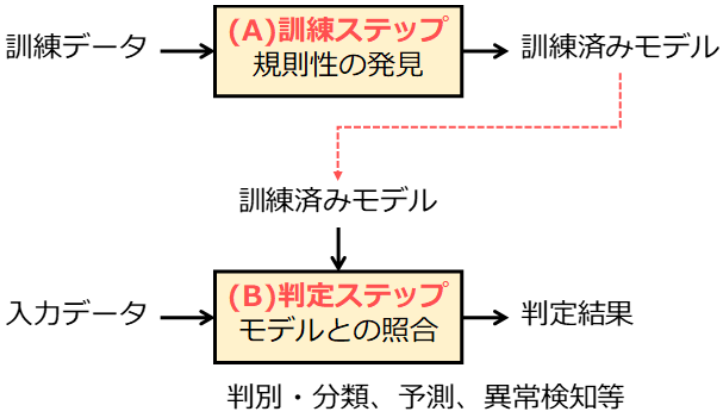


図２　機械学習の処理ステップ

図3は、訓練データから規則性が得られ、それが判定に利用されるイメージを示したものである。上段では空間的な規則性が判別・分類に利用され、下段では時間的な規則性が予測や異常検知に利用される例である。

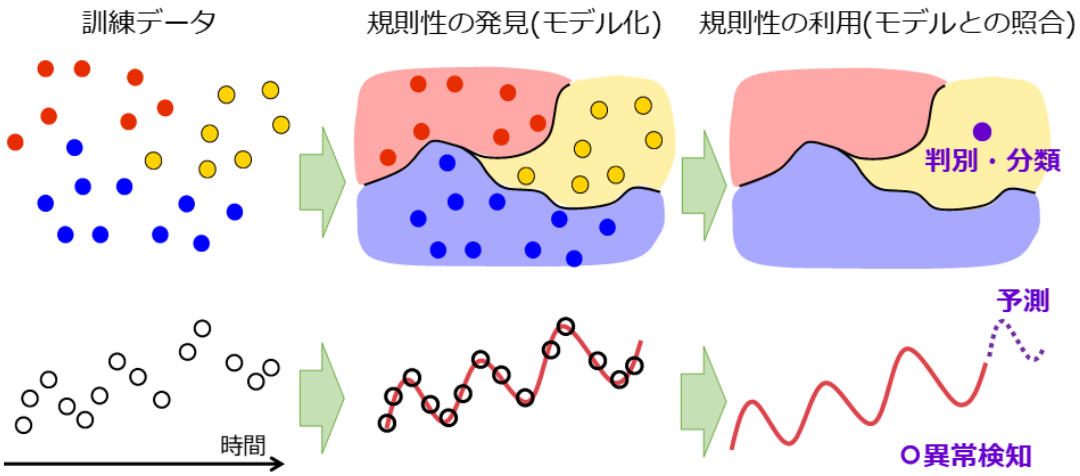


図３　機械学習における規則性の発見と利用（出典［5］）

(A) は学習ステップ、(B) は推論ステップと呼ばれることもあるが、本稿ではそれぞれ訓練ステップ、判定ステップと呼ぶ。「学習」という言葉はコンピューターの自発性を感じさせるが、実際は人間が「訓練」するステップである。同様の理由からデータやモデルについても「学習データ」「学習済みモデル」ではなく「訓練データ」「訓練済みモデル」[2]と呼ぶ。また、「推論」というと高次の処理がイメージされるが、(B) は訓練済みモデルとの照合による判定である。本稿では「記号推論」と呼ぶやや高次の処理にも触れることから、混同を避けるためにも、(B) には「推論」ではなく「判定」という言葉を用いる。

2．研究開発を実施する意義

2－1．現状認識および問題点

2－1－1　機械学習技術の進化とインパクト

現在、第 3 次 AI ブームと言われるが、これを牽引しているのは、深層学習をはじめとする機械学習技術である[5]。機械学習技術は、いわば「前例に倣う AI」であり、多数の前例をデータとして与えると、その中に潜む規則性を見出し、新たに与えられたデータに対して、前例に倣った判断（データの分類・予測・異常検知等）ができるようになる。

特に、従来に比べて規模の大きい多層ニューラルネットワーク（Deep Neural Network、図 4）を用いた機械学習技術である深層学習技術は、様々な特定用途で人間を上回る精度を達成して注目を浴びている。図 5 は、共通の大規模画像データセットを用いた一般物体認識コンペティション ILSVRC（The ImageNet Large Scale Visual Recognition Challenge）のトップチーム成績の推移を示したものである。2012 年に深層学習によって認識エラーが一気に 10% 減るという衝撃的な性能改善が示され、その後 3 年で人間の精度を上回るようになった。

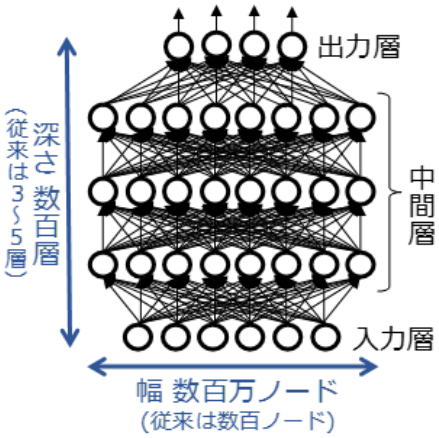


図 4　多層ニューラルネットワーク

深層学習において、前例データから得られた規則性は、多層ニューラルネットワーク中のリンクの重みという形で記録され、人間にとって明示的な規則として理解することが困難なもの（いわゆるブラックボックス）である。深層学習は、人間が形式知化すること難しいレベルの判断規則まで自動獲得し得るということが、人間を上回る精度の達成を可能にしているものと考えられる。

図 5　ILSVRC トップチームのエラー率の推移
（http://www.image-net.org/challenges/LSVRC/ をもとに作成）

2－1－2　システム開発方法のパラダイム転換

深層学習をはじめとする機械学習技術は、研究コミュニティーや新聞・雑誌メディアを賑わせているだけでなく、実際に様々なシステムに埋め込まれ、実社会での応用・実用化が急速に広がっている。

機械学習技術を用いると、システムの動作を大量データから帰納的に定義することができる。この方法は、プログラムコードを書いてシステムの動作を定義する従来の（演繹的な）開発スタイルと大きく異なる。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

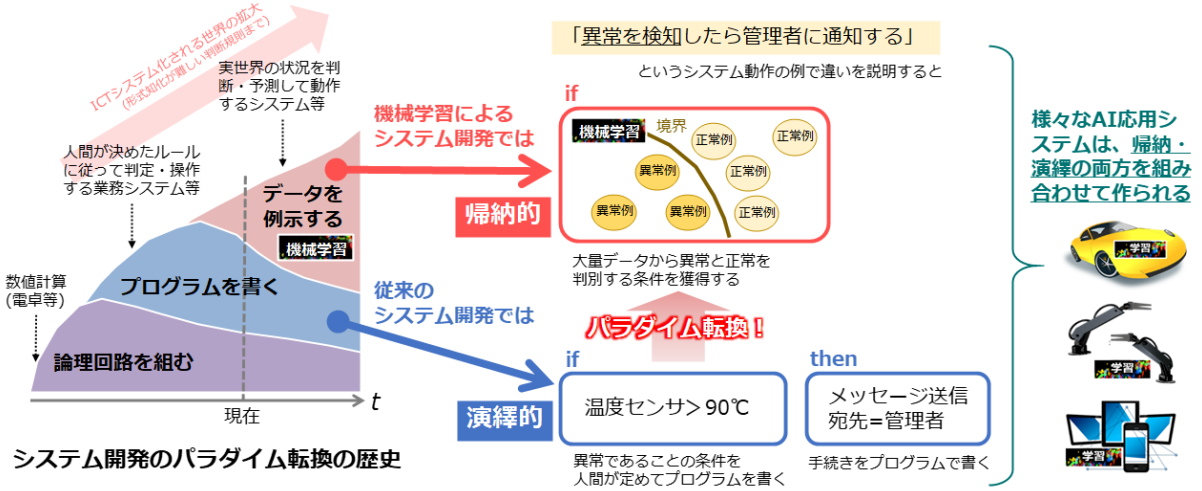


図６　システム開発のパラダイム転換

方法論や技術は必ずしも使えず、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[1][2][3][4]のである（図 6）。

図 6 を用いて、このパラダイム転換について説明する。図 6 の左部に示したように、システム開発の方法は、古くは「論理回路を組む」という方法から始まり、「プログラムを書く」という方法が生まれ、さらに「データを例示する」という新しい方法が生まれてきた。それにつれて、ICT システム化される対象世界が広がってきた。この流れの中で「データを例示する」という機械学習によるシステム開発は、それまでの演繹的な考え方から帰納的な考え方への転換と見ることができる。図 6 の中央に簡単な例を示したが、システムに「異常を検知する」という動作をさせたいとき、従来の演繹的開発では、例えば「温度センサ＞ 90℃」のように異常であることの条件を人間が決めてプログラムを書くのに対して、機械学習を用いた帰納的開発では、人間が異常例・正常例のデータを与えると、システムが異常・正常の判別条件を自動的に獲得する。なお、システム開発のすべてが「データを例示する」方法になるわけではなく、「論理回路を組む」「プログラムを書く」「データを例示する」という 3 通りの方法を組み合わせて作られていくことになる。

２－１－３　対応遅れが招く社会的・産業的ダメージ

このようなシステム開発のパラダイム転換が起きていることに対して、対応が遅れると、日本が社会的・産業的に大きなダメージを被ると懸念される。

（1）産業的な問題：人材スキルや方法論の刷新

「プログラムを書く」という（演繹的な）システム開発では、従来のソフトウェア工学に基づく方法論や技術が整備されてきた。しかし、「データを例示する」という帰納的なシステム開発では、従来の方法論や技術が必ずしも通用しなくなる。それらが道具として使えないケースが増えるだけでなく、それらの方法論や技術を習得してきた人材も活躍できる場面が減ってくる。代わりに、帰納的なシステム開発に有効な方法論・技術や、それらを習得した新しいスキルの人材が必要になる。このような人材スキルや方法論の刷新に追従できないと、日本のソフトウェア開発・システムインテグレーター等の産業が競争力を失いかねない。

以下では、従来の方法論や技術が必ずしも通用しなくなるということを説明する。

よく知られている従来の方法論としてV字モデル（図7）がある。V字モデルに基づくシステム開発では、要求定義→基本設計→詳細設計→コーディングというステップで開発し（V字の左半分）、単体テスト→結合テスト→システムテストというステップで検証する（V字の右半分）。この方法論は、システムの機能（仕様）が、抽象度の高いレベル（上位）から具体的なレベル（下位）へと階層的に分解できることを想定している。そして、システム全体の品質は、下位から上位への積み上げによって確保できることを想定している。しかし、機械学習を用いてデータの例示によってシステムの動作を定義する場合、例示するデータセットにデータを1つ加えるだけで、その影響が全体に及ぶ。このCACE（Changing Anything Changes Everything）[6]と呼ばれる性質のため、機械学習を用いた帰納的なシステム開発にV字モデルはそぐわない。

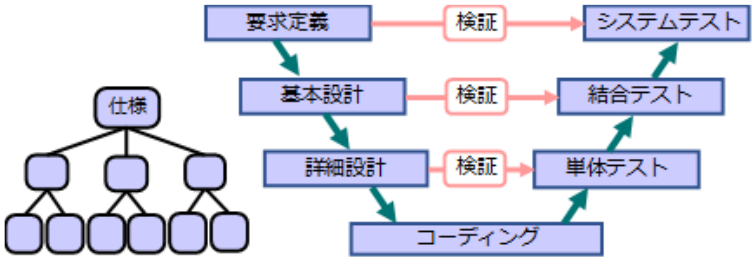


図7　従来の開発方法論：V字モデル

図8は、従来型の開発方法論・プロセス（図8の青色のフロー）に沿って、機械学習型の開発を進めた場合（図8の赤色のフロー）に発生する主な課題を列挙したものである[7]。そのいくつかについて具体例を述べる。

テストに関する課題の1つは「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未整備」というものである。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がかりがない。例えば自動運転におけ

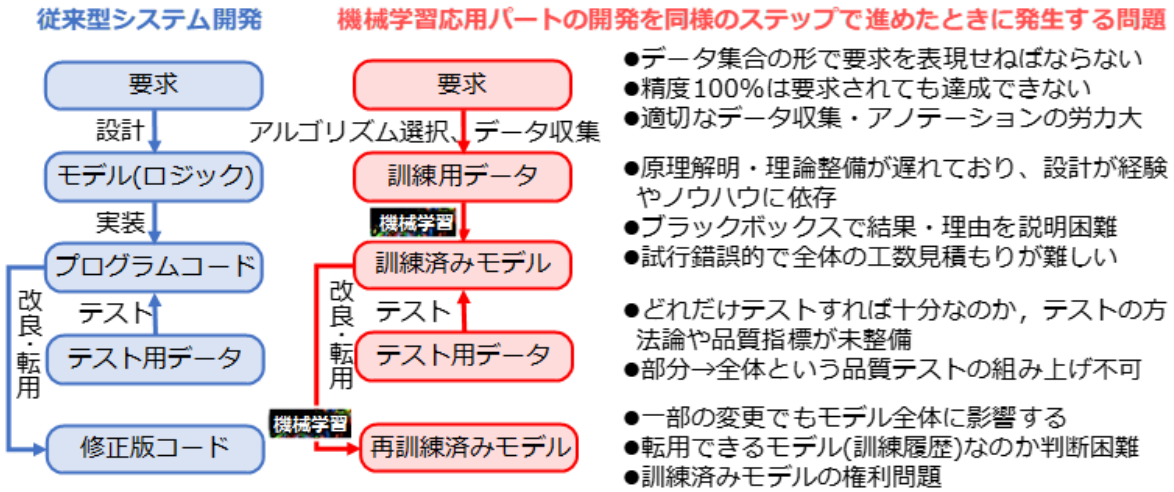


図8　従来の開発方法論をAI応用システム開発に適用したときに発生する課題

る環境認識という応用では、「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」「マンホールから人が顔を出すかもしれない」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。

また、深層学習等、機械学習の方式によっては「ブラックボックスで結果・理由を説明困難」という課題がある。この課題は特にシステムが事故を起こしたときに問題になる。従来型の場合、前述のようにV字モデルに従って階層的に設計とテストが行われているので、事故が起きたときは、その現象に基づいて階層的に該当パートを絞り込んでいくことで、事故の原因解明を進めることができる。しかし、ブラックボックス型の機械学習方式を使っていると、起きた現象・事故に対して、どのような理由でそれが起きたのかを説明することが難しい。しかも、前述したCACEと呼ばれる性質（訓練のために与えるデータを1つ変えるだけで全体の動作が大きく変わり得る）が、原因解明をいっそう厄介なものにしている。したがって、機械学習を説明可能なものにしたり、機械学習型と従来型のソフトウェアの組み合わせによってブラックボックスの振る舞いをモニタリングしたり、事故発生時の原因解明につながる技術が求められる。

（2）社会的な問題：AI応用システムの安全性に不安

（1）で述べたように、機械学習技術を用いたAI応用システム開発では、従来型の開発方法論・技術が必ずしも有効でなく、システムの品質保証・動作保証に問題が残る[1]。品質保証・動作保証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の社会実装が急激に進むと、そこで発生した問題や事故が思わぬ社会問題につながりかねない。動作保証ができず安全性に問題があると、製造物責任法（PL法）やリコール制度の対象になるリスクさえも生じる。以下、そのような懸念が顕在化したいくつかの事例をあげる。

- Adversarial Examples攻撃：

  深層学習が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）を加えたときに、深層学習がそれまでと全く異なる誤認識結果を出すことがある。このような加工画像（Adversarial Examples）を生成して、深層学習に誤認識をさせる攻撃はAdversarial Examples攻撃と呼ばれている。道路標識を対象とした実験で、深層学習が停止標識と認識できていた画像に、人間には気にかけない程度のごく小さなノイズを乗せて、速度制限標識と誤認識させることができたと報告されている[8]。もし自動運転の道路標識認識システムにこの攻撃をかけたとしたら、自動運転車両は停止標識で止まらずに走行し、事故を引き起こす恐れがある。

- Microsoft Tay公開停止：

  Microsoftが開発した機械学習型チャットボットTayが、2016年3月23日にインターネット上に公開されて、わずか一日で公開を停止するという事件があった[9]。Tayは、悪意のあるユーザーとの対話から言葉を獲得した結果、差別と陰謀論に染まってしまい、不適切な発言を繰り返すようになってしまった。

1　機械学習型の方法論・技術が未整備の現状では安全性・品質等に不安があるという意味であり、従来型に比べて、機械学習型の安全性・品質が本質的に低いというわけではない。例えば、機械学習アルゴリズムそのもののプログラムはコードが短く、判定ステップの応答も高速だという性質は、バグが入り込みにくく、リアルタイム処理性能も安定し、安全性に寄与する。機械学習型は従来型と性質が異なるので、異なる対応・考え方を整備していくことが必要なのである。

- 訓練データに潜む差別・偏見：
  悪意を持った攻撃だけでなく、知らずに差別・偏見を取り込んでしまうケースもある。例えば、罪を犯した人間の再犯率を予測するために、人間による判断結果を用いて訓練したとする。もしその判断を行った人間が特定人種に偏見を持っていたならば、機械学習による判断結果もその傾向を反映してしまう。また、個々の訓練データに偏見を持った判断があるか否かだけでなく、訓練データの分布の偏りや量が差別を生むケースも指摘されている。例えば､特定人種について誤認識が多いことが問題になった事例[10]や、採用判定で性別による差別が起きてしまった事例[11]等が報告されている。

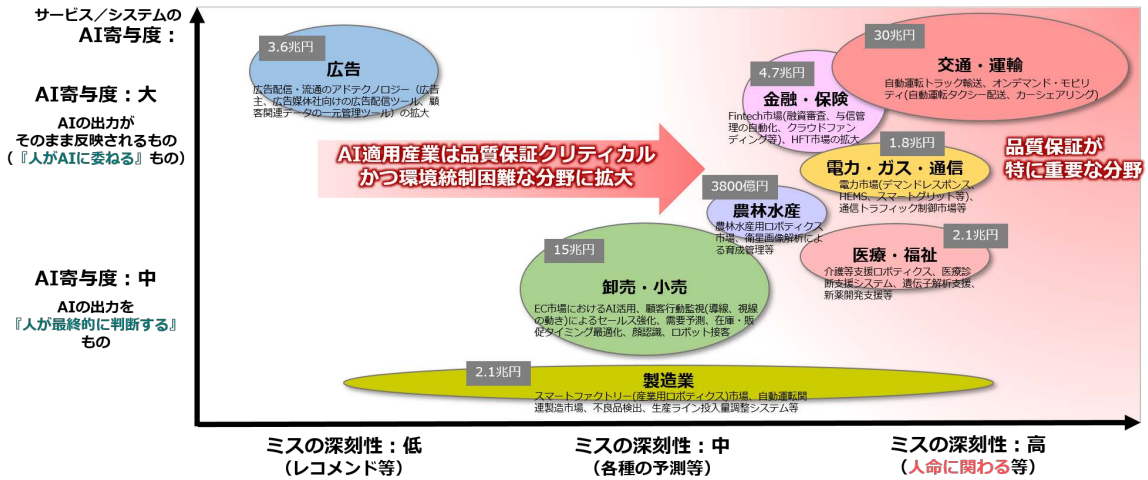
２－１－４　国内外の状況・動向

（１）AI 適用産業における品質保証の重要性

AI･機械学習の応用は様々な産業に広がっているが、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の深刻性は応用毎に異なる。そして、その深刻性があまり高くない分野から、人命に関わるような深刻性の高い分野へまで、応用が広がってきている（図 9）。

AI･機械学習の応用と品質保証の関係性を考える上で、その応用の 3 つの性質、すなわち「ミスの深刻性」「AI 寄与度」「環境統制困難性」に着目する。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なものになる。

- ミスの深刻性：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：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：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。



金額は 2030 年の AI 適用産業の予想市場規模（EY 総合研究所「人工知能が経営にもたらす創造と破壊」[12] を参考）

図９　産業分野と品質保証クリティカル性

図 9 には、ミスの深刻性を横軸に取り、AI 寄与度を縦軸に取って、産業分野をマッピングした。厳密には、個々の応用によって性質が異なるので、産業分野を一塊でマッピングするのはやや強引だが、おおまかな傾向は表わせている。図 9 では軸としていない環境統制困難性も含めて考えると、時間的な推移として、実用化が早かったのは、ミスの深刻性と環境統制困難性の低い応用だったが、最近はミスの深刻性と環境統制困難性の高い応用も実用化が試みられるようになってきた。

例えば機械学習技術を用いた応用として、ショップサイトにおける商品のレコメンド機能や､郵便番号や宛名住所の文字認識による郵便物の自動振り分け（郵便区分機）は 10 年以上前から実用化されている。これらの応用は、ミスの深刻性は低く（ユーザーの好みに合わないレコメンドがなされたら無視される／ショップへの印象は低下するものの)、環境統制困難性も低い（ショップサイト上の機能としてコンテキストはコントロールされている)。しかも、日々大量のデータが得られ、ミスの深刻性も低いので、事前よりもむしろ結果としての性能・品質が重視される。それに対して、期待が高まっている自動運転では、ミスの深刻性が高く（人命に関わる)、環境統制困難性も高い（車が走行する道路環境では天候変化や急な飛び出し等の様々な現象が起こり得る)。そのため､事前の品質保証が極めて重要なものになる。

（2）業界・研究コミュニティーの動向

このように AI･機械学習の応用システムが社会･産業に広がっていき、その品質保証の重要性が急速に高まっている。しかし、AI･機械学習技術そのものの研究開発が活発なのに比べて、それらの技術を組み込んだシステムの安全性・信頼性の確保や品質保証に関する研究開発については、まだ取り組みが十分でない。２－１－２節で述べたように、システム開発のパラダイム転換が起きていることに対して、安全性・信頼性の確保や品質保証のために、新たな理論・原理から研究開発に取り組んでいくことが必要になっている。

このような認識は業界・研究コミュニティーでも急速に高まってきている。特に最近 1～2 年はその動きが顕在化してきた。図 10 には、産業界、AI 研究コミュニティー、ソフトウェア工学研究コミュニティーでの最近の動向を簡単にまとめた（詳細情報は付録 2 を参照)。

まず、産業界では、問題意識が急速に高まっているが、根本的な解決策はまだ見えていない。その中でも、自動運転分野が具体的な検討で先行している。また、AI の品質･安全性に関する標準化の議論も始まっている。

これに対して、AI 技術そのものの研究コミュニティーは、機械学習の精度・性能を高める競争が激しく、開発法への関心はやや低い。ただし、社会における AI という視点からの議論が活性化しており､AI の安全性や説明可能性も含めた社会的な課題やあるべき姿の検討が行われており、これらは開発法そのものの研究ではないが、開発法を設計する上での重要な指針になり得るものである。

一方で、ソフトウェア工学研究コミュニティーでは、最近 1～2 年で問題意識の共有や研究会・ワークショップ等での具体的な議論が立ち上がっている。国内では 2017 年初め頃からシンポジウムでの招待講演、ワークショップや企画セッション等が多数実施され、議論が活性化してきた。特に 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 が発足し、5 月のキックオフシンポジウムには 500 名を超える参加者が殺到し

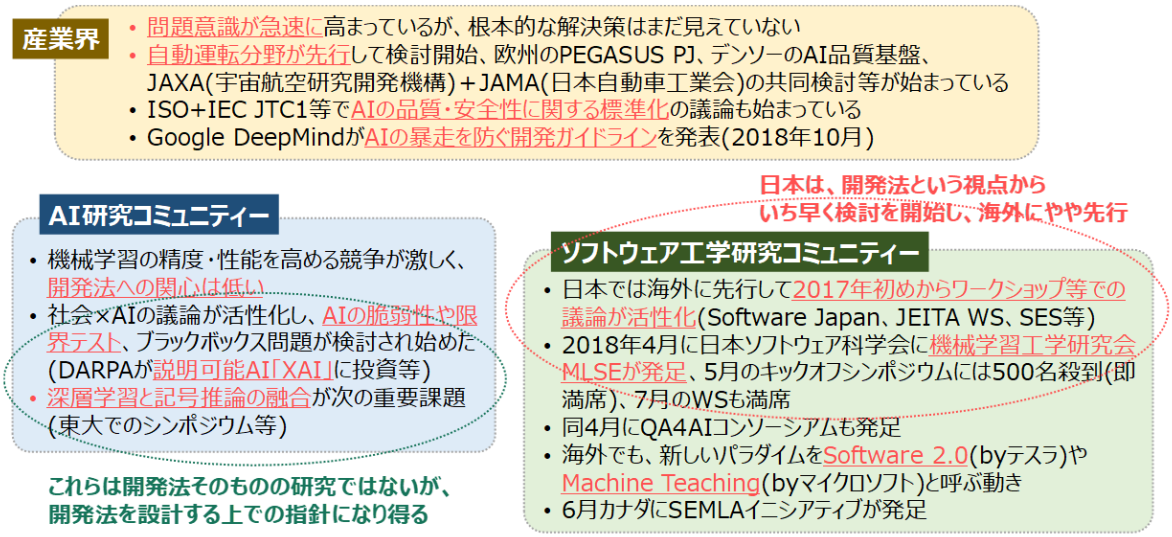


図 10　業界・研究コミュニティーの動向

たことは、この問題に対する関心の高さが顕在化した出来事であった。また、同じく 4 月に AI プロダクト品質保証コンソーシアム QA4AI も発足した。海外でも、新しいパラダイムを Software 2.0 や Machine Teaching と呼ぶ動きがあり、国際学会での招待講演として Software 2.0 の話題が取り上げられる等、同様にホットなトピックになりつつある。このような海外の動向に対して、日本は、開発法や品質保証という視点からいち早く検討を開始し、海外にやや先行していると思われる。

（3）わが国の AI 戦略・政策から見ると

AI（人工知能）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。わが国でも、総理指示を受けた AI 研究の体制として「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､AI 技術の研究開発が強化されている。一方、内閣府に「人間中心の AI 社会原則検討会議」が設けられ、AI の社会的・経済的・倫理的・法的な課題（Ethical, Legal and Social Issues：ELSI）を含む社会から見た AI への要請が検討されている[2]。

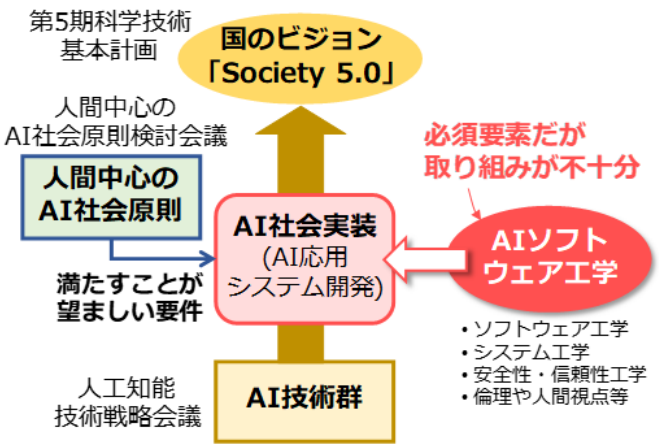


図 11　国の AI 関連政策における位置付け

第 5 期科学技術基本計画に示されたわが国のビジョン「Society 5.0」に向けて、AI 技術の社会実装は重要な役割を担う。その際、AI 技術そのものの研究開発を推進する「人工知能技術戦略会議」と、人間中心の AI 社会原則を提言する「人間中心の

2　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」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」等で検討されてきた。特に「AI ネットワーク社会推進会議」では、透明性、利用者支援、制御可能性、セキュリティ確保、安全保護、プライバシー保護、倫理、アカウンタビリティという 8 つの原則を掲げた「AI 開発原則（案）」や、「AI 利活用原則（案）」等を示してきた。内閣府は、2018 年 4 月に「人間中心の AI 社会原則検討会議」を発足させ、先行の検討結果を踏まえつつ、同年度末までに「人間中心の AI 社会原則（案）」をまとめる予定である。

AI 社会原則検討会議」という 2 つの政策ではまだ足りない。ここで必要になるのは、「人間中心の AI 社会原則」が示す姿（満たすことが望ましい要件）を実現するために、AI 技術群を用いて、どのように AI 応用システムを作ればよいかというソフトウェア工学的な開発方法論･技術体系である。本プロポーザルでは、それを「AI ソフトウェア工学」と呼んでいるが、その取り組みはまだ十分でなく、「AI ソフトウェア工学」の研究開発を推進するための政策が必要である[3]（図 11）。

２－２．社会・経済的効果

AI ソフトウェア工学の研究開発を強化する必要性について述べてきたが、その社会･経済的効果を以下にまとめる。

第一に、社会に急速に広がっていく AI 応用システムに関して、その安全性・信頼性の確保につながる。わが国が目指す Society 5.0（超スマート社会）においても、あるいは、いままさに第 3 次 AI ブームの中で次々に生み出されている新サービスにおいても、AI 応用システムの安全性・信頼性は強く求められる。しかし、AI 応用システムの安全性・信頼性を確保するための開発方法論・技術体系は現状まだ整備・確立されておらず、AI ソフトウェア工学の推進によってそれが可能になる。

第二に、AI・機械学習の抱える品質リスクが引き起こしかねない事故・社会問題を回避することにつながる。２－１－３節で述べたように、AI 応用システムの品質保証･動作保証のための方法論が確立されていないことで､大事故や社会問題を引き起こす懸念がある。AI ソフトウェア工学の研究開発の推進によって、このリスク･懸念を回避することが可能になる。

第三に、AI 品質を日本の強みにして、AI 適用産業の国際競争力を強化することにつながる。AI 応用システムの品質基準･安全性基準が定められていない場合、安価だが粗悪な AI 商品･サービスが他国から国内市場に流れ込むかもしれない。また、単なる価格競争に陥り、品質・安全性に問題のある商品・サービスが市場にあふれて事故が頻発するかもしれない。そのような事故が多発するようであれば、AI 応用システムに対する懸念が高まり、厳しい規制や国内 AI 関連産業の委縮を招くものと考えられる。そのような事態を回避し、市場の健全化と産業の活性化を促進するためには、国として適切な品質基準・安全性基準を定めることが求められる。そのとき、その基準が十分な根拠を持ったものであることや、その基準を満たすようなシステム開発のための方法論が確立されていることが重要であり、国際競争力を生む。その裏付けとなる新しい理論・技術体系が AI ソフトウェア工学の研究開発から生み出され得る。

２－３．科学技術上の効果

科学技術上の効果としては、繰り返しになるが、システム開発のパラダイム転換に対応

3　JST CRDS では、2017 年 9 月からこの「AI ソフトウェア工学」に関する戦略プロポーザルの検討を進めてきたが、この発行までの間に国の政策に動きがあった。2018 年 6 月 15 日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内閣に設置され､その第 1 回が 7 月 27 日、第 2 回が 9 月 28 日に開催された。7 月 27 日の会議を受けて、イノベーション政策強化推進のための有識者会議「AI 戦略実行会議」が設置され、9 月 28 日に「AI 戦略（案）」が提言された。その有識者ペーパー「AI 戦略（案）全体俯瞰図」では「AI 工学」があげられた。この「AI 工学」は「AI ソフトウェア工学」を含むターゲットと解釈できる。
https://www.kantei.go.jp/jp/singi/tougou-innovation/dai2/siryo1.pdf

する新しい学問分野「AI ソフトウェア工学」の創出ということが大きい（その名称については国内では「機械学習工学」、海外では「Software 2.0」等が使われており、必ずしも「AI ソフトウェア工学」という名称で分野が築かれるとは限らないが）。

この AI ソフトウェア工学は、従来のソフトウェア工学の技術的な拡張になる。その拡張が技術的にどのような価値を持つかというと、従来のソフトウェア工学が想定しているのはクローズドな世界であるのに対して、AI ソフトウェア工学はオープンな世界へと、そこで扱えるシステムタイプを拡大できる（図12）。

また、この新しい学問分野は、ソフトウェア工学に AI 技術を融合したものとなるが、さらに、安全工学やシステム工学の考え方も融合したものになる。そして、これからの産業界に不可欠な、このような新しいスキルを持った技術者の育成を促進することにもつながる。

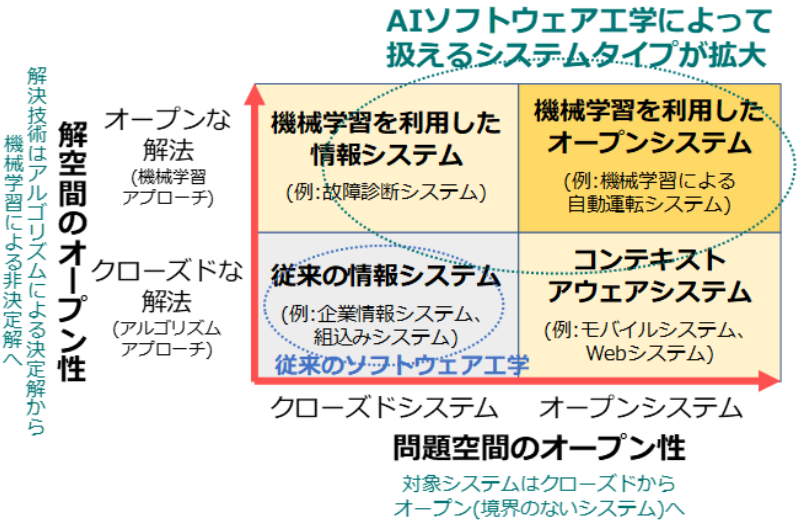


図 12　扱えるシステムタイプの拡大

（青山幹雄教授（南山大学）の資料［13］をもとに作成）

3．具体的な研究開発課題

本プロポーザルで提唱する AI ソフトウェア工学は、従来のソフトウェア工学に対し、新しい世代のソフトウェア工学として位置付けている（p.1 の図 1）。このような新しい学問分野を構築・確立していくための研究開発課題を 2 つの面からあげる。1 つは AI ソフトウェア工学の体系化であり、もう 1 つは基礎研究として重要な技術チャレンジである。

AI ソフトウェア工学の体系化の全体像（現状でのおおまかなイメージ）について 3 － 1 節で説明し、基礎研究として重要な技術チャレンジの内容について 3 － 2 節で説明する。AI ソフトウェア工学の全体像を踏まえ、基礎研究として重要な技術チャレンジをどのようにして抽出したか、そのプロセスについては付録 2 の（3）に記載した。

3－1．AI ソフトウェア工学の体系化

システムの安全性・信頼性の確保は一つの技術で解決するものではない。総合的な取り組みが不可欠であり、そのためには、全体像を押さえつつ、バランスよく、見通しよく、技術開発を進めるべきである。

その際、従来のソフトウェア工学は学問分野として歴史があり、体系化も進んでいるので、AI ソフトウェア工学の体系を考えていく際のベースにすることができる。具体的には、IEEE Computer Society によってまとめられたソフトウェア工学の知識体系 SWEBOK（Software Engineering Body of Knowledge、https://www.computer.org/web/swebok ）を

表 1　ソフトウェア工学の体系の拡張

SWEBOK V3.0 の知識領域	機械学習型システム開発を含めた場合に必要になる拡張 (MLSEBOK)
ソフトウェア要求	機械学習方式に依存した要求の厳密性、機械学習精度に応じた要求 (人間の介在程度等) の見直し
ソフトウェア設計	従来型と機械学習型の役割分担設計、要求を表現する訓練データ設計、データ収集・加工手順の設計
ソフトウェア構築	機械学習プロセスの実行と進行判断、機械学習実行の計算リソースの確保・最適化
ソフトウェアテスティング	機械学習パートのテスト方法、テスト用データ設計、機械学習の CACE 性 (Change Anything Changes Everything)、プロセス品質とプロダクト品質
ソフトウェア保守	初期訓練時からの環境変化の監視と対策、追加訓練・再訓練、継続的なデータ収集
ソフトウェア構成管理	機械学習方式とパラメータの管理、訓練データ・訓練履歴・訓練済みモデルの管理、訓練済みモデルの再利用性
ソフトウェアエンジニアリングマネージメント	従来型と機械学習型の混在マネージメント、工数見積もり (格段に難しい) の方法論
ソフトウェアエンジニアリングプロセス	トライアル (PoC) ステップを含む機械学習型システム開発に特徴的なプロセスやフロー、機械学習 API・専用言語
ソフトウェアエンジニアリングモデルおよび方法	従来型 (演繹・トップダウン) と機械学習型 (帰納・ボトムアップ) の組み合わせモデル・使い分け指針、様々な機械学習方式の特性・適性
ソフトウェア品質	機械学習の確率的動作 (100% 保証は不可)、機械学習を含むシステムの品質保証、要求に対する定量指標 (多様性・曖昧性)
ソフトウェアエンジニアリング専門技術者実践規律	機械学習の公平性・説明責任・透明性・倫理 (FATE)、機械学習に関わるデータとモデルの知的財産権
ソフトウェアエンジニアリング経済学	トライアル (PoC) やデータに関わる経済視点
計算基礎	機械学習アルゴリズム、GPU プログラミング
数学基礎	統計数理
エンジニアリング基礎	データ工学

参考にすることができる。表１の左列に SWEBOK V3.0 の知識領域を示した。

表１では、さらに右側の列に、機械学習型システム開発を含めた場合に、従来のソフトウェア工学の考え方・手法から、さらにどのような点で拡張が必要になりそうかを列挙した。このような検討を進めて拡張した方法論･技術体系が、AI ソフトウェア工学の全体像につながっていくものと考えられる。

国内では日本ソフトウェア科学会の機械学習工学研究会 MLSE でこのような議論がなされているが、同研究会による機械学習工学ワークショップ MLSE2018 では、SWEBOK に相当する知識体系を機械学習工学についてもまとめていこうという提言が出され、その知識体系の仮称として「MLSEBOK：Machine Learning Systems（or Software）Engineering Body of Knowledge」と呼ぶ案が示された（表１は JST CRDS からのポジションペーパー[7]で示したものだが、同種の提言は MLSE2018 の中で複数あった）。

なお、ソフトウェア工学を中心に置いたが、AI 応用システムの安全性･信頼性を確保するためには、ソフトウェア系（従来型＋機械学習型）だけでなく、ハードウェア系も含めたシステム全体を対象とし、開発プロセス、機能安全、品質保証を総合的に扱う必要がある。したがって、技術分野としては、狭い意味でのソフトウェア工学にとどまらず、広く安全工学やシステム工学もスコープに含めて全体像をとらえていくべきである。そういう意味でも、表１のような整理は、AI ソフトウェア工学の体系化の検討の出発点や見通しを得る粗い地図にはなるが、真の体系化はより踏み込んだ検討・議論が必要である。

３－２．基礎研究として重要な技術チャレンジ

付録２の（３）に詳細を示したが、AI ソフトウェア工学の全体像を踏まえて、考えられる技術課題を（i）機械学習自体と（ii）全体システムという 2 タイプに分け、さらに、それぞれに対して、（A）実践から得られる知見の体系化のアプローチと（B）理論・原理に関わる学術研究のアプローチという 2 つの面から整理した。（A）については産業界主体で取り組みが進むに違いないが、（B）については基礎研究としての取り組みを強化する施策が望まれる。そこで、（B）の取り組みが必要あるいは特に有効なものを、技術的にある程度まとまった塊で捉え、基礎研究として重要な 4 つの技術チャレンジとして抽出した（図 13）。

以下では、それら基礎研究として重要な 4 つの技術チャレンジ（①機械学習自体の品質保証、②全体システムとしての安全性確保、③ブラックボックス問題への対策、④問題を効率的に解く工学的な枠組み）について､それぞれの概要と現状の技術シーズを説明する。

なお､AI ソフトウェア工学の体系全体としては､前出の表１や付録１の表２に示したように、システム開発プロセス全般に及び、要求分析、構築・運用・保守等のプロセス、工数見積もりや予算管理等まで含む幅広いものである。それらも含む総合的な技術開発・方法論整備が必要な学問分野となるが、図 13 にも示したように、産業界を中心とした（A）実践から得られる知見の体系化のアプローチが進められており、（B）理論･原理に関わる学術研究のアプローチが重要になる技術課題は、安全性・信頼性に関わるものが多くなった。

３－２－１　機械学習自体の品質保証

第一の技術チャレンジは「データ集合に着目した機械学習の品質（精度、公平性等）の

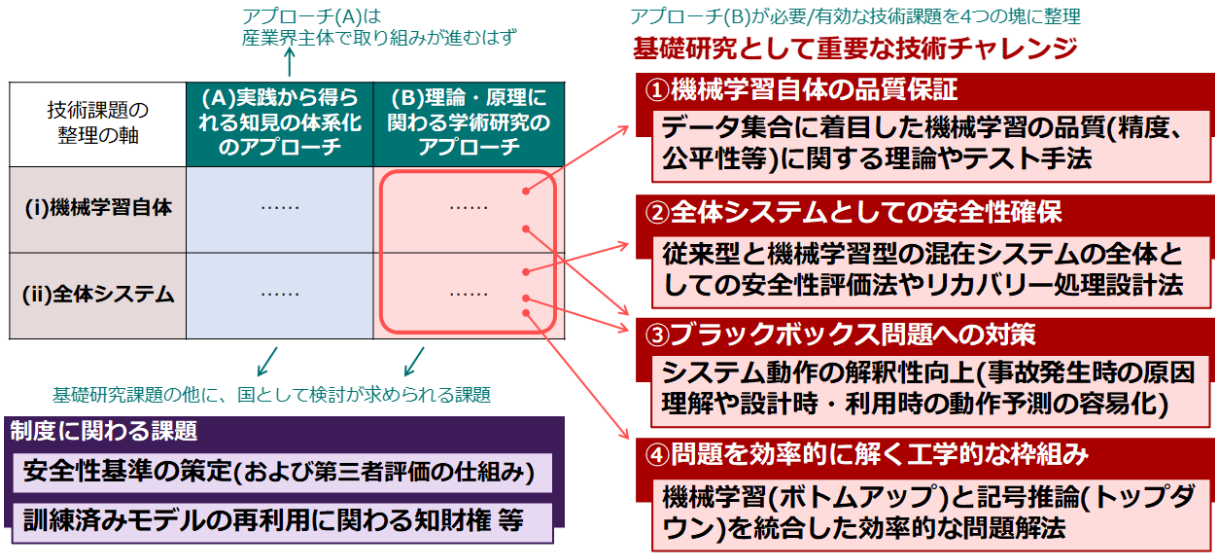


図 13　基礎研究として重要な技術チャレンジの抽出

理論やテスト手法」である。機械学習の精度や想定外リスクを左右するのは訓練データ集合なので、データ集合の良さに着目して評価するアプローチが検討されている。また、結果の公平性や訓練済みモデルの再利用性も、データ集合の良さから考えることが有効と考えられる。

【技術シーズ】

機械学習のテストのカバレッジやパターン量を増やすための様々な手法が提案されている。国内では国立情報学研究所（National Institute of Informatics：NII）等で先行的に取り組まれているが（NII の中島震・石川冬樹ほか）、理論的整備や公平性確保等、まだ多くの研究課題が残されている。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等のアイデアが知られている[14][15][16]。

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、それを拡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（図 14）。

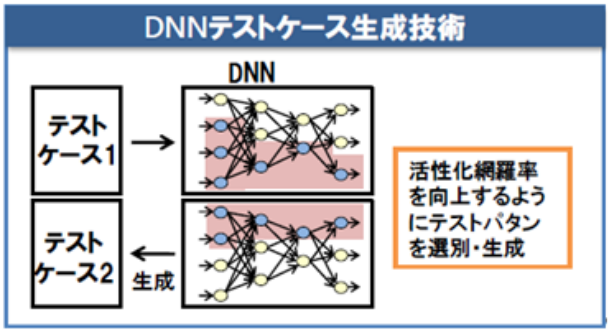


図 14　ニューロンカバレッジ（出典［16］）

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。

機械学習における公平性の問題は、「公平性配慮型データマイニング」（Fairness-Aware Data Mining）や「機械学習における公平性・説明責任・透明性（FAT）」（Fairness,

Accountability, and Transparency in Machine Learning）等の名称で、4 ～ 5 年前から国際学会のワークショップやチュートリアルでも取り上げられるようになり、国際的に研究が活発になっている。国内でこの研究に取り組んでいる神嶌敏弘（産業技術総合研究所）によると[17]、この問題は不公平さの検出と不公平の防止という 2 つのタスクを中心に取り組まれている。また、訓練に用いるデータ集合に差別・偏見が潜んでいないかという視点だけでなく、訓練時にセンシティブな属性をどう扱うかという視点も必要になる。人種・性別等の属性を除外して機械学習を実行したとしても、他の属性に人種・性別等と相関の高いものがあれば、不公平な結果を出してしまうかもしれない。また、通常、機械学習の精度と公平性の確保はトレードオフ関係が生じ、応用・目的に応じた設計が必要である。

3－2－2　全体システムとしての安全性確保

第二の技術チャレンジは「従来型と機械学習型の混在システムの全体としての安全性評価法やリカバリー処理設計法」である。機械学習型コンポーネントが 100% 保証はできないことを想定した全体システム設計や、機械学習型の入力・出力をモニタリングして、例外処理やリカバリー処理・多重化等を組み込むことが有効である。

【技術シーズ】

機械学習による判断に、従来法による安全性保証を組み合わせる Safe Learning という考え方がある。図 15 の例では、従来法によって、演繹的に安全が確保できる範囲を決定し、機械学習を用いてアクションを選択する際には、この安全が確保された範囲（Safety Envelope）内に限定する[18]。これによって、機械学習を組み込んだとしても、安全性が保障できる。国内では、JST ERATO MMSD（メタ数理システムデザイン）プロジェクト（研究総括は NII の蓮尾一郎）等で取り組まれている。

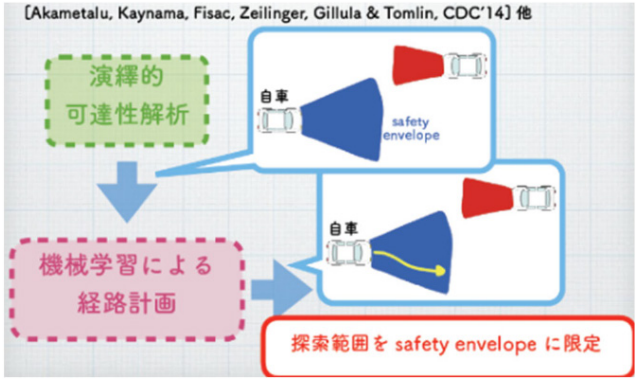


図 15　Safe Learning（出典［18］）

また、機械学習も 1 つのコンポーネントとして扱い、コンポーネント間の相互作用から安全性分析・ハザード解析手法である STAMP/STPA（Systems-Theoretic Accident Model and Processes / System-Theoretic Process Analysis）や FRAM（Functional Resonance Analysis Method）がある[19][20]。その際に判断の確信度に応じてリスクの少ない方に倒すアイデア等が検討されている（宇宙航空研究開発機構 JAXA と日本自動車工業会 JAMA の共同研究[21]、デンソーの AI 品質基盤[22]等）。

米国国防高等研究計画局（Defense Advanced Research Projects Agency：DARPA）では「Assured Autonomy」[32]という研究開発投資プログラムが実施されている。自動運転車やドローン等の自律システムの安全性保証をターゲットとして、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）を起動するような枠組みが研究開発されている。

全体システムとしての安全性確保のためのテストに関しては、３－２－１でも触れたようなカバレッジを広げることや、そもそもまれにしか発生しない想定外ケースでもテストすること等が課題になる。この課題に対して、シミュレーションによって多様なケースを生成してテストするというアプローチがとられている。一例として、Waymo（Google の自動運転会社）の取り組みを紹介する。Waymo は公道走行距離の実績において、既に他社と比べて圧倒的といえる 1000 万マイルを達成している。そして、この公道走行だけでなく、様々な状況をシミュレーションで生成した仮想空間内の走行距離は 70 億マイルに到達しているとのことである。シミュレーションを併用することで、実世界ではめったに起こり得ない特殊なケースも試すことが可能になる。

シミュレーションによるテストに関連して、NEC －産総研 人工知能連携研究室から発表された希少事象発見技術[33]がある。この技術では、まれにしか発生しない不具合･高リスクの条件を、シミュレーションと機械学習を組み合わせることで、短時間で見つけることができ、効率の良い不具合の発見・修正を可能にする。

３－２－３　ブラックボックス問題への対策

第三の技術チャレンジは「システム動作の解釈性向上（事故発生時の原因理解や設計時･利用時の動作予測の容易化）」である。機械学習応用に不安を与えている要因として､機械学習が判定したとき､その理由を人間が理解できないというブラックボックス問題[4]がある[23]。(すべての機械学習方式がブラックボックスというわけではないが)システムにブラックボックスな部分があると、設計時や利用時に、設計者あるいは利用者がシステムの動作を予測することが難しく、システムの安全性を確保する上で妨げになる。また、もしシステムが事故を起こしてしまった場合、その原因理解や責任の切り分けが難しいということにもなる。さらに、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（第 22 条）が盛り込まれたため、GDPR 規制遵守という面でもブラックボックス問題への対策が必要になっている[24]。

機械学習には様々な方式があり､すべての機械学習方式がブラックボックスというわけではない。例えば､決定木や線形回帰等の方式では､訓練結果の規則性（ルールやモデル）が人間にも理解しやすい形式(if-then ルールや線形和式等）で得られるので、ブラックボックス問題を回避できる(ホワイトボックス型と言われる)。一方、昨今、注目されている深層学習をはじめとするニューラルネットワーク系の機械学習方式は､訓練の結果がネットワーク構造中の重み等に反映されるため､人間がその意味を直接解釈することは難しく、典型的なブラックボックス型である。

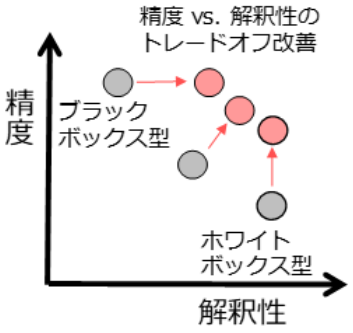


図 16　機械学習の精度と解釈性のトレードオフとその改善方向

4　ブラックボックスという言葉は、その中で起きていることがまったくわからない状態をイメージさせる。しかし、機械学習に関してブラックボックスという場合、その中で起きている信号の流れは観測でき、動作のメカニズムもプログラムコードとして記述されているので、上記のイメージでとらえるのは適切ではない。ここでは、その中で起きていることを人間が解釈することが難しいという意味で、ブラックボックスという言葉を用いている。

機械学習方式の精度と解釈性はおおむねトレードオフ関係になる。上で述べた方式であれば、深層学習はブラックボックス型で解釈性が低いが、精度が高く、決定木や線形回帰はホワイトボックス型で解釈性が高いが、精度が低い。

そこで、機械学習における精度と解釈性の両立が求められている（図 16）。今日、ブラックボックス型だが高い精度が得られる深層学習が広く活用されるようになったことから、精度と解釈性の両立に向けた取り組みとして、ブラックボックス型に解釈性を与えるという取り組みが盛んになってきた。また、機械学習自体の解釈性を高めるという点だけでなく、機械学習技術が組み込まれた AI 応用システム全体としてシステム動作の解釈性が求められるという点も考えておくべきである。その意味で、ユーザインタフェースも重要な要素になる。システムの振る舞いが人間に理解しやすいことや、人間とシステムの間の認知ギャップを避けること等も、AI 応用システムのユーザインタフェース設計技術として検討されていくべきである。

【技術シーズ】

AI・機械学習の解釈性を高めるための研究開発として、米国 DARPA がファンディングする「説明可能な AI」（eXplainable Artificial Intelligence：XAI）プロジェクト[25]が有名である（コラム③参照）。DARPA の XAI 自体の目的は、システムの安全性よりも軍用途の意思決定支援の意味合いが強いように思われるが、XAI という言葉は産業界・学術界に広がり、安全性の面からも重要な研究課題として認知され、技術開発が活発になってきている。

機械学習に解釈性を与える（機械学習の精度と解釈性を両立させる）XAI のアプローチは、次の 3 通りに整理される（図 17）。

- Deep Explanation：深層学習の状態解析によるアテンションヒートマップや自然言語説明生成等
- Interpretable Models：もともと解釈性の高いモデルを用いた機械学習（ホワイトボックス型の機械学習の精度を向上させる）
- Model Induction：ブラックボックス型の機械学習の振る舞いを近似する解釈性の高いモデルを外付けで作る

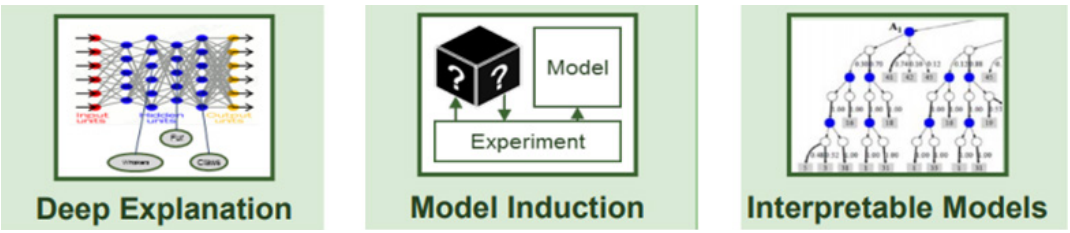


図 17　XAI の 3 つのアプローチ（出典［25］）

DARPA の XAI プロジェクトにおいても、これら 3 通りのアプローチが並行して推進されている。Deep Explanation は、特に深層学習を用いた画像認識システムにおいて、盛んに取り組まれている。例えば、認識結果を出した根拠として、画像中のどの部分に着目したのかを示すアテンションヒートマップを示すといった方法である。

【コラム③】DARPA の XAI プロジェクト

米国 DARPA は、2017 年～ 2021 年の 5 年間を実施期間とし、説明可能な AI（XAI）に関する研究開発投資プログラムを推進している。DARPA XAI は、今後現れるであろう AI、特に機械学習に基づく“パートナー”を人間の兵士が理解し、信頼し、管理することを目指したプロジェクトであり、高度な学習機能を維持しながら従来よりも説明可能なモデルを生成する機械学習技術を開発している。具体的には、DARPA のミッションに対応する 2 つのタスク、Data Analytics（機密情報分析）と Autonomy（自律システム）を設定している。Data Analytics タスクは技術的にはマルチメディアデータの分類問題であり、映像からターゲットを自動的に識別する際に人間の分析官に対して識別根拠を説明する（例えば、敵と識別した根拠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など自律システムの強化学習問題であり、自動パイロットを行う際に人間の操作者に対して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。ここで説明とは、システムの出力や特徴を人間が解釈できる意味情報に繋げることを意味しており、本文でも述べるように Deep Explanation（深層学習においてどの特徴がどの部分の識別に効いているのかを示す）、Interpretable Models（Random Forest や Bayesian network、Probabilistic Logic においてネットワーク内のノードの意味や相関関係を示す）、Model Induction（モデルをブラックボックスとして入出力をより簡単で解析可能なモデルで再現する）等の手法が検討されている。前 2 者が内包的、後の 1 つが外延的な説明といえるだろう。同時に、ユーザーが理解可能な形に翻訳するインタフェースの開発もプロジェクトに含まれている。さらに、既存の心理学的な定理を拡張し、説明の効果を予測するため計算可能な“説明の定理”を開発することもプロジェクトに含まれている。これらインタフェース技術と心理学の融合によって理解可能な説明を提示できるとされている。一方で、ユーザモデリングやパーソナライズ、インタラクティブ機械学習、可視化技術などは含まないことが明示されている。プロジェクトの評価に関しては、精度と説明可能性にトレードオフがあることを明言しており、両方を測定するとしている。なお、説明性は心理学的にアプローチされているため、説明の効果測定ではその明快さや有用性に関するユーザーの満足度をレーティングするとされている。また、DARPA XAI では AI システムの判断理由（説明）を知ることができれば正しく動いているかどうかは人間が判断できるという考え方か、品質保証という言い方は出てこない。現在、説明可能モデルと説明インタフェースには 11 チーム（UC Berkley、CMU、Xerox PARC 等）が、説明の心理学には IHMC（The Institute for Human & Machine Cognition）の 1 チームが取り組んでいる。チームあたりの予算規模は年間 $800K ～ $2M である。特に、UC Berkley のチームは深層学習を用いて自動運転のための XAI に取り組んでおり、自動運転時における AI 判断の説明をヒートマップで視覚的に示す研究を進めている。予定では、本稿執筆時点（2018 年 10 月）に Phase1: Technology Demonstration の評価 Eval1 が行われており、年明けにも結果が公表されることが期待される。さらに、Phase2: Comparative Evaluation からは各チームが独創的な問題を設定し、2019 年秋、2020 年秋にそれぞれ Eval2、Eval3 が予定されている。最終成果物はオープンにされる予定である。

国内でも取り組まれており、NEC が 2012 年から事業化している異種混合学習は、決定木的な場合分けに重回帰分析式を組み合わせたモデルをデータから自動生成することで、解釈性の高いモデル（Interpretable Models のアプローチ）でありながら、高い精度を達成している。また、富士通は 2018 年に、深層学習 Deep Tensor にナレッジグラフを組み合わせた Model Induction のアプローチによる取り組みを発表した[26]。

３－２－４　問題を効率よく解く工学的な枠組み

第四の技術チャレンジは「機械学習（ボトムアップ）と記号推論（トップダウン）を統合した効率的な問題解法」である。深層学習が広がったことで帰納型・ボトムアップ型の問題解決が主流になっているが、本来、演繹型・トップダウン型の記号推論も重要である。それらのあるべき組み合わせや新しい問題解決の枠組みを探究することは、AI 応用システムのアーキテクチャー設計という面で重要である。

この第四の技術チャレンジは、前述した他の技術チャレンジ、特に第二・第三の技術チャレンジとやや重なる面がある。第二の技術チャレンジは安全性、第三の技術チャレンジは解釈性を高めることを狙っており、その狙いを達成するために、機械学習（ボトムアップ）と記号推論（トップダウン）を統合するというのも一つの解法になり得るからである。しかし、機械学習（ボトムアップ）と記号推論（トップダウン）を統合することで得られる効果は、安全性や解釈性の向上だけとは限らない。より広い視点から AI 応用システムのアーキテクチャーの革新を図ることは、基礎研究として重要である。

【技術シーズ】

人間のトッププロ棋士を破って衝撃を与えた AI 応用システム「AlphaGo」（Google DeepMind）[27]も、この一例である。問題を木の探索問題（トップダウン）として定式化し、どの枝を優先的に探索するかの決定を深層学習（ボトムアップ）の結果から得ている。

また、国内の事例として、新エネルギー・産業技術総合開発機構（New Energy and Industrial Technology Development Organization：NEDO）のプログラム「次世代人工知能・ロボット中核技術開発」で、九州工業大学・産業技術総合研究所・理化学研究所が取り組んでいる「データ駆動型人工知能と論理知識型人工知能の融合による解釈可能な自動運転システムに関する研究」（図 18）においても、トップダウンとボトムアップの統合による効率的な解法が試みられている[28]。ここでは、センサ等の入力から状況を判定するデータ駆動型 AI がボトムアップ処理（機械学習）で、知識（オントロジー）に基づいて、その状況で起こり得る危険を予測する論理知識型 AI がトップダウン処理（記号推論）である。2 つが連携することで、効率よく適切にアクションを決定できるようになる。

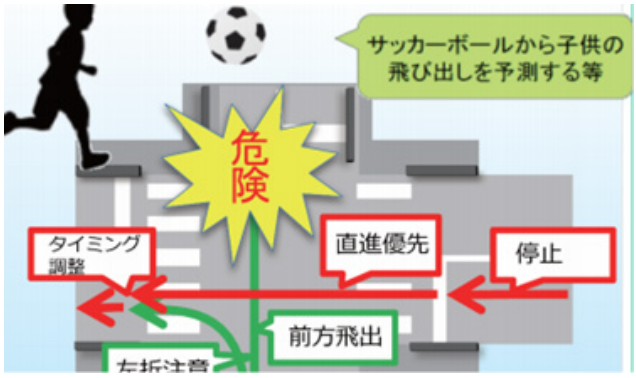


図 18　九工大・産総研・理研による「データ駆動型人工知能と論理知識型人工知能の融合による解釈可能な自動運転システムに関する研究」（出典 [28]）

【コラム④】もう一つの課題：擬人化インタフェース設計

本プロポーザルでは、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フェース（擬人化インタフェース）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。そのような認知ギャップを最小化するような設計手法が求められる。

4．研究開発の推進方法および時間軸

4－1．施策の設計方針

AI ソフトウェア工学の研究開発を推進するためには、新たな施策を講じていくことが必要になる。効果的な施策の設計･実行のためには、以下のような方針･考え方が鍵となる。

設計方針
1．問題解決型の基礎研究をトップダウンに加速
2．人工知能×ソフトウェア工学の技術融合を促進し、新しい技術分野を創出、事業競争力を強化
3．①学術研究＆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
4．競争が始まっている自動運転分野で先行実証し、他分野へ展開
5．日本の先行優位性を活かす迅速な施策推進

以下、これら 5 つの設計方針を順に説明する。

（1）問題解決型の基礎研究をトップダウンに加速

AI ソフトウェア工学の研究開発を推進するには、問題解決型の基礎研究をトップダウンに加速するファンディングプログラムが必要である。既存のトップダウン的なファンディングプログラムは基本的にシーズプッシュ型、科研費（科学研究費助成事業）はボトムアップ的であり、この目的には必ずしも十分なスキームではない。

図 19 に（a）シーズプッシュ型と（b）問題解決型の違いを基礎研究という視点で示した。

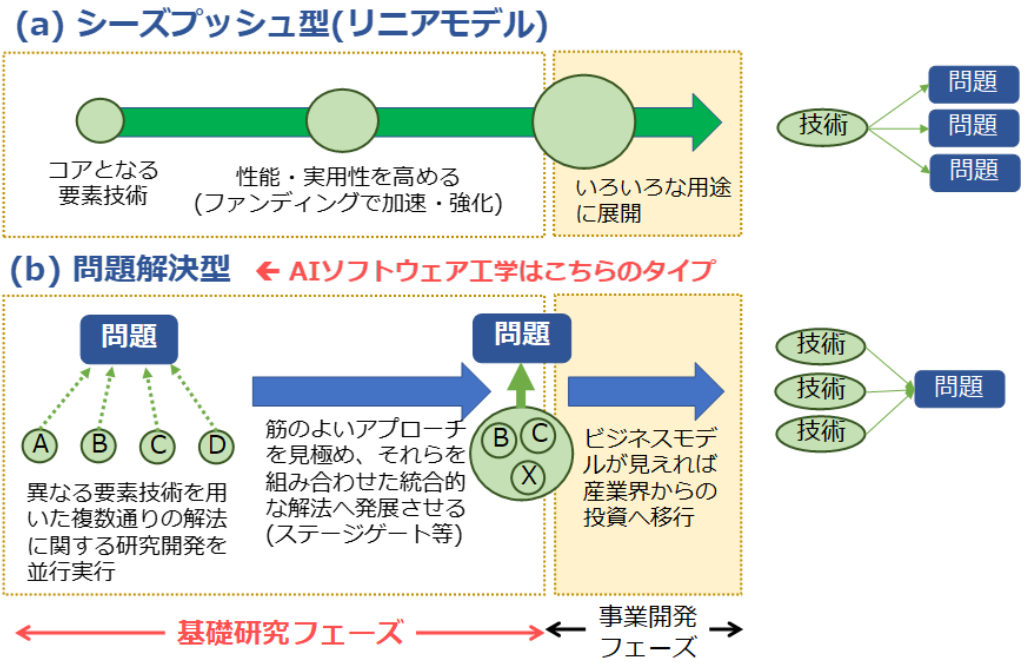


図 19　シーズプッシュ型 vs 問題解決型

（ａ）シーズプッシュ型は旧来のリニアモデルに沿ったものであり､コアとなる要素技術（例：5G 通信、量子コンピューター、○○型デバイス等）があり、その性能・実用性を高めていくフェーズとして基礎研究が位置付けられる。そのフェーズを加速・強化するためのファンディングプログラムが用意されている。事業開発・産業化のステップでは、その要素技術がいろいろな用途に展開される。

一方、（ｂ）問題解決型の基礎研究は、解決すべき問題（例：一般画像認識、機械翻訳、異常検知等)が先にあり､問題に適した解法を見出すまでが基礎研究である。そのフェーズでは、異なる要素技術を用いた複数通りの解法が検討され、それらの研究開発が並行して進められる。その中から筋のよいアプローチを見極め、それらを組み合わせた統合的な解法に発展させるところまでが基礎研究である。解法が見出されれば、その解法・技術の実用化には産業界が競争して投資する。ファンディングプログラムとしては、問題を明確に定義するが、解くための技術・アプローチは限定しない。ステージゲート型のファンディングも活用する。

（ａ）（ｂ）のどちらが産業的に主導権を取れるかは､技術や問題の性質によるが､近年､情報系産業においては（ｂ）が主流､つまり､最終価値に対価を払う傾向が強まっている。

AI ソフトウェア工学も（ｂ）タイプの基礎研究が必要であり､これを国としてトップダウンに推進するファンディングプログラムが求められる。

（２）人工知能×ソフトウェア工学の技術融合を促進

人工知能（AI）×ソフトウェア工学（SE）の技術融合を促進し、新しい技術分野を創出、事業競争力を強化する。

この方針の背景として、国内学術界についてやや見受けられる次のような課題がある(すべての大学研究が該当するわけではなく、これらの課題にまったく該当しないケースは多々あるが)。まず､実データに基づく研究の一層強化が必要である。特に AI 分野の革新技術は実データに基づく研究から生まれることが多い。また、ソフトウェア開発力が軽視されがちというのも懸念点である。AI 分野ではデータとソフト開発力が競争力を生む。大学における学術研究とはいえ、ソフトウェア開発力は重要である。さらに、大学におけるソフトウェア工学研究が産業界の開発現場からともすると乖離しているという指摘が産業界からしばしば聞かれる。システム開発のパラダイム転換を通して産学関係・業界の変革を進めるような意識を持つ必要があるかもしれない。

このような弱点を克服し、AI ソフトウェア工学における真の競争力を生み出すためには､人工知能（AI）×ソフトウェア工学（SE）の技術融合を促進し､新しい技術分野の創出、事業競争力の強化に結びつけていくことを強力に推し進めることが望まれる。その具体的な施策化においては、次の（３）で述べるような体制造りを産学連携で進めることが必要である。

（３）３つの活動を密連携させて推進

AI ソフトウェア工学の研究開発は、①学術研究＆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る。

「①学術研究＆人材育成」は主に大学等の学術研究機関が担うが、それが､「②実応用での技術実証」や「③基準策定＆標準化」と結びつく形で取り組まれることが望ましい。そ

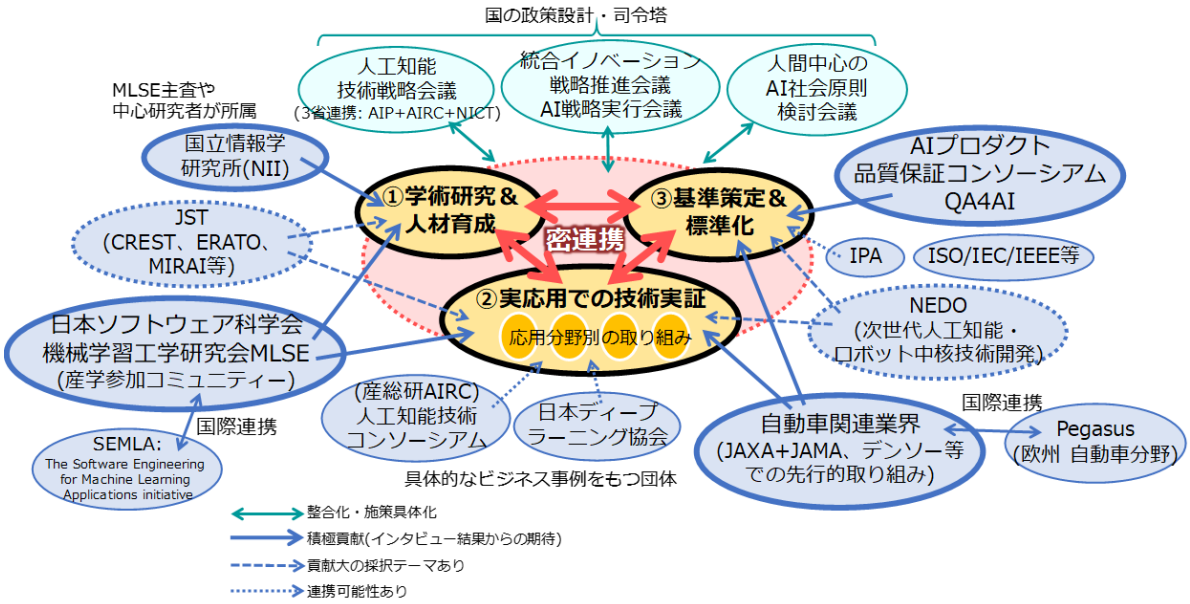


図 20　中核となる 3 つの活動と関連組織

のためには、産学連携のプロジェクトチーム体制を作ったり、密連携を推進する司令塔組織を設置したりするのが有効である。実際、図 20 に示すように、中核となる 3 つの活動（上記の①②③）に関わる様々な活動や団体･組織が立ち上がってきており、それらを束ねて推進する体制・司令塔の早期設置が望まれる。

（4）競争が始まっている自動運転分野で先行実証し、他分野へ展開

AI ソフトウェア工学の研究開発では、産業分野毎に要求される AI 品質が異なることを踏まえて取り組むことが重要である。その一例として、AI プロダクト品質保証コンソーシアム QA4AI では、自動運転、工場（製造）、スマートスピーカー、クリエイティブ（3D 動画生成）という 4 つのワーキンググループを設置して検討を進めている。

その中でも特に自動運転を最重点産業分野ととらえる。自動運転は、将来に向けて国際的にも競争が始まっている重要分野である。この自動運転分野 AI ソフトウェア工学を先行実証し、そこで生み出した AI 品質の国際競争力を、産業特性を考慮しつつ他分野に展開するという進め方がよい。自動運転分野は品質保証クリティカルで、業界の危機意識も強く、世界的に検討が先行している。ここでの成果が他分野にも波及する可能性が高い。

参考のため、自動運転分野における品質保証への取り組み事例を以下に示す。

- JAXA + JAMA：AI 時代のソフト品質保証の考え方について共同検討（2018 年 3 月プレスリリース）。
- デンソー：自動運転を含む AI 搭載システムの品質保証体系（仕組み）の構築を進めるとともに、1. アプリケーション、2. サブシステム、3. コンポーネント（モデル／データ）の 3 階層で品質技術を開発している。
- ERATO 蓮尾メタ数理システムデザインプロジェクト：数学を基礎とした形式手法を自動運転システム開発に適用。
- ペガサス（Pegasus）プロジェクト：ドイツ BMWi（経済エネルギー省）主導で自動車関連主要企業が参加、自動運転の試験方法の標準化を進める。

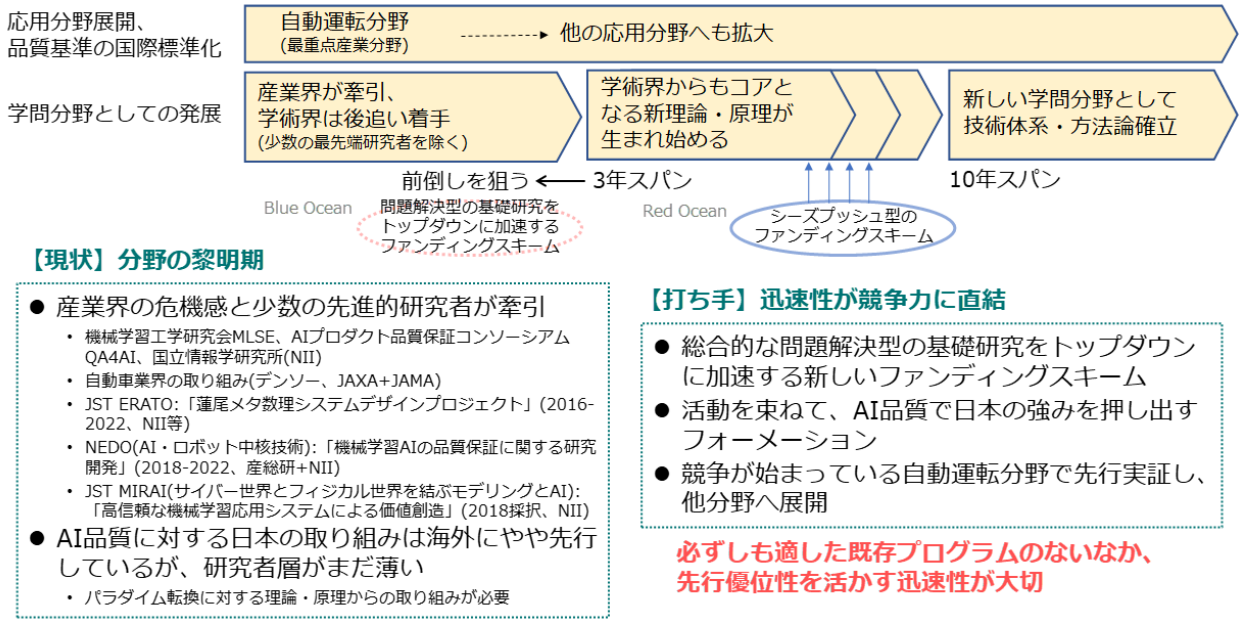


図 21　研究開発推進の時間軸

自動運転分野は品質保証クリティカルな産業分野として、AI ソフトウェア工学の取り組みの必要性が高いことを述べたが、機械学習技術は自動運転に限らず、今後開発される様々なシステムに広く組み込まれていくに違いない。いまの AI 技術の中核が機械学習であることに加えて、IoT（Internet of Things）や CPS（Cyber-Physical Systems）と呼ばれる実世界接点を持った応用システム分野においても機械学習は不可欠な技術になりつつある。IoT・CPS に関わる産業分野も今後ますます成長すると見込まれている。実世界の状況に応じてきめ細かに動作するシステムをプログラミングする方法として、実世界で起こりえる状況の膨大な可能性を数え上げて、それぞれのケースに対する動作を事細かに（演繹的に）定義することは現実的ではない。事例に基づく（帰納的な）機械学習型のシステム開発の方が相性が良い。自動運転ほど品質保証クリティカルでなくとも、安全で信頼性の高い機械学習応用システムを効率よく開発する方法論･技術体系として、AI ソフトウェア工学が必要な応用分野は大きく広がっていくはずである。

（5）日本の先行優位性を活かす迅速な施策推進

AI ソフトウェア工学は、学問分野としてまだ黎明期だが、AI に関わる他分野と同様に、動きがはやいことは間違いない。しかも、現状、国内での取り組みは海外にやや先行した面もあり、決して後追い状況ではなく、世界的にリードできるチャンスがある。また、品質基準・安全性基準等を打ち出して、標準化活動を進めるにあたっても先行優位性はキープしたい。したがって、日本の先行優位性を活かす迅速な施策推進が強く望まれる。ただし、開発で先行し、品質の高さでも優位性を達成したとしても、日本だけのガラパゴス企画になってしまうと、産業面で必ずしも得策ではない。国際的な協力関係・グローバル技術を意識した取り組みを進めることも肝要である。

AI ソフトウェア工学の全体としては、産業界が牽引するものと思えるが、パラダイム転換に対応する新しい理論・原理の創出や、学問分野としての技術体系・方法論の確立（10年かけて取り組むものである）には、学術的な基礎研究が不可欠である。

図 21 に AI ソフトウェア工学の研究開発推進の時間軸のイメージを示した。ここでは、迅速性を確保しつつ、中長期的に学問分野として発展させる仕掛けに課題がある。シーズプッシュ型のファンディングスキームは､ある種､研究開発のレッドオーシャン時期になって投資するようなものになっており、むしろ、ブルーオーシャン時期に投資し、先行優位性や研究者層の厚みを確保するために活用されるべきと考える。

４－２．国の施策の役割

２－２節や２－３節で述べたように､本プロポーザルは､国が目指す社会（Society 5.0）において、AI 応用システムの安全性･信頼性を確保するためのものである。そして、産業においても､AI 品質による国際競争力の確立を狙う。それを支える学術研究として、新世代のソフトウェア工学「AI ソフトウェア工学」を新しい学問分野･技術体系として確立していくことを目指す。

このために国の施策の役割を、産業界・学術界との関係とともに、図 22 にまとめた。

３－２節で述べたように､AI ソフトウェア工学の技術課題に対しては、産業界での取り組みを中心とした（A）実践から得られる知見の体系化のアプローチと、学術界での取り組みを中心とした（B）理論・原理に関わる学術研究のアプローチの両面から解いていくことが必要である。産業界は、（A）だけでは問題を解決できず、AI 品質による国際競争力を確立するためには、AI 品質を裏付ける学術研究に期待している。

ただし、この学術研究の姿は、AI を中心とした情報系の分野において、理論だけではなく、ソフトウェアとしての実装や、実データでの検証・分析が一体化した、問題解決型の研究である。旧スタイルからの変貌が求められる。そのため､４－１節の（３）でも述べたように､学術研究と実応用での技術実証を密連携させた研究フォーメーションが望ましい。

また、日本において、産業界が、危機感や国際競争を背景として、AI ソフトウェア工学への取り組みを推進しているのに対して、学術界は、まだ少数の先端研究者が取り組みを進めている段階で、研究者層を厚くしていかねば、現在の先行優位性を活かしていくことが難しくなる。

このような状況において、国の施策が重要なポイントは、まず、密連携した活動の司令塔としての役割である。その下で、安全性・品質基準を策定・実施していくことも重要な国の施策の役割である。さらに、学術界に対して、理論・ソフトウェア・実データを統合した研究への変貌と、問題解決型の基礎研究の推進を促すことである。

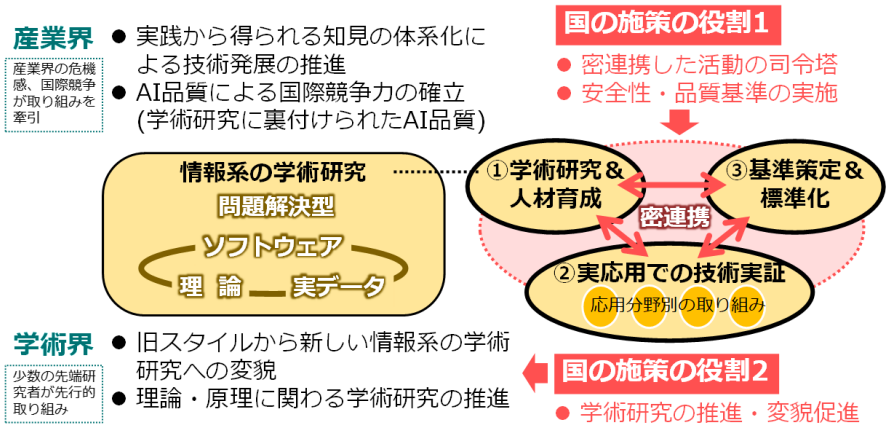


図 22　国の施策の役割

付録１．検討の経緯

国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（CRDS）では、2018年度の戦略スコープ策定委員会において、戦略プロポーザルを作成すべきテーマの候補として、本テーマを選定し、検討チームを発足させた。本検討チーム自体は2018年5月から活動を開始したが、それに先立ち、CRDSのシステム・情報科学技術ユニットでは2017年9月から予備調査も実施していた。その間、有識者インタビューやワークショップも開催し、本テーマの研究開発状況の把握や研究課題・方向性の議論を深めてきた。

以下に、有識者インタビューとワークショップの概要を含め、検討の経緯を記す。

（1）有識者インタビュー

本プロポーザルの作成にあたり、研究内容や推進体制、研究シーズ等について意見をうかがうため、関連する研究領域に高い専門性を有する有識者への個別インタビューやワークショップを実施した。2017年9月～2018年9月に実施し、以下の方々からお話を聞いた（50音順、敬称略、所属・役職はインタビュー実施当時、複数回インタビューした方もあり）。

青山幹雄	南山大学 理工学部 ソフトウェア工学科 教授
石井正悟	情報処理推進機構（IPA）社会基盤センター イノベーション推進部 調査役
石川冬樹	国立情報学研究所（NII）アーキテクチャ科学研究系 准教授
石濱直樹	宇宙航空研究開発機構（JAXA）研究開発部門 第三ユニット 主任研究開発員
岡崎直観	東京工業大学 情報理工学院 情報工学系 教授
奥村洋	株式会社デンソー 東京支社 電子基盤先行開発室 室長
門田暁人	岡山大学大学院 自然科学研究科 産業創成工学専攻 教授
桑島洋	株式会社デンソー 東京支社 電子基盤先行開発室 担当係長
権藤正樹	イーソル株式会社 取締役 CTO 兼 技術本部長
佐藤洋行	株式会社ブレインパッド マーケティングプラットフォーム本部 副本部長
菅原啓介	宇宙航空研究開発機（JAXA）研究開発部門 第三ユニット 研究開発員
関口智嗣	産業技術総合研究所（AIST）理事 兼 情報・人間工学領域 領域長
谷幹也	日本電気株式会社（NEC）セキュリティ研究所 所長
中江俊博	デンソー 東京支社 電子基盤先行開発室 開発３課長
中島震	情報・システム研究機構 戦略企画本部 兼 国立情報学研究所（NII）情報社会相関研究系 教授
中島秀之	東京大学大学院 情報理工学系研究科 特任教授、札幌市立大学 学長
中田登志之	東京大学大学院 情報理工学系研究科 教授
西康晴	電気通信大学大学院 情報理工学研究科 情報学専攻 講師
橋本真太郎	宇宙航空研究開発機（JAXA）研究開発部門 セキュリティ・情報化推進部 情報化基盤課 主事補

蓮尾一郎　国立情報学研究所（NII）アーキテクチャ科学研究系 准教授
東中竜一郎　日本電信電話株式会社（NTT）メディアインテリジェンス研究所 主任研究員、メディアインテリジェンス研究所 兼 コミュニケーション科学基礎研究所 NTT 上席特別研究員
樋口知之　統計数理研究所 所長
福住伸一　理化学研究所 革新知能統合研究センター 社会における人工知能研究グループ 科学技術と社会チーム 研究員
藤吉弘亘　中部大学 工学部 ロボット理工学科・情報工学科 兼 大学院工学研究科 情報工学専攻・ロボット理工学専攻 教授
松尾豊　東京大学 工学部 システム創成学科 特任准教授
丸山宏　株式会社 Preferred Networks 最高戦略責任者
水野修　京都工芸繊維大学 情報工学・人間科学系 教授
明神智之　宇宙航空研究開発機（JAXA）研究開発部門 第三ユニット 研究開発員
向山輝　情報処理推進機構（IPA）研究員
本橋洋介　日本電気株式会社（NEC）AI・アナリティクス事業開発本部 シニアデータアナリスト 兼 データサイエンス研究所
本村陽一　産業技術総合研究所（AIST）人工知能研究センター 首席研究員
鷲尾隆　大阪大学 産業科学研究所 教授

（2）俯瞰ワークショップ

本プロポーザルの作成にあたり、下記の通り、俯瞰ワークショップを開催した。その詳細についてはワークショップ報告書 CRDS-FY2017-WR-11 を発行済みである。
http://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-11.html

「機械学習型システム開発へのパラダイム転換」
【開催日時】2017 年 12 月 8 日（金）13:30 ～ 18:00
【開催場所】TKP 市ヶ谷カンファレンスセンター　カンファレンスルーム 7C

【プログラム】
総合司会：茂木強（JST 研究開発戦略センター フェロー）
13:30 ～ 13:40　開催挨拶
木村康則（JST 研究開発戦略センター 上席フェロー）
13:40 ～ 14:00　開催趣旨説明
福島俊一（JST 研究開発戦略センター フェロー）
14:00 ～ 14:30　機械学習工学に向けて
丸山宏（株式会社 Preferred Networks 最高戦略責任者）
14:30 ～ 15:00　ソフトウェア工学×機械学習
～ディペンダブルな機械学習ソフトウェア・システム開発に向けて～
中島震（国立情報学研究所 情報社会相関研究系 教授）
15:10 ～ 15:40　AI を活用した社会価値創造の取り組みと今後の課題
谷幹也（日本電気株式会社 セキュリティ研究所 所長）

15:40～16:10　機械学習型システム開発へのパラダイム転換 －MPRGの研究紹介―
藤吉弘亘（中部大学 工学部 ロボット理工学科 教授）
16:10～16:40　機械学習型システム開発のビジネス的課題
佐藤洋行（株式会社ブレインパッド マーケティング
プラットフォーム本部 副本部長 チーフデータサイエンティスト）
16:50～17:05　未来社会創造事業の次年度公募テーマに向けた問題意識
前田章（JST 未来社会創造事業「超スマート社会」領域総括）
17:05～17:50　総合討論
ファシリテーター：福島俊一（JST 研究開発戦略センター フェロー）
17:50～18:00　まとめ・閉会挨拶

（3）AI ソフトウェア工学の技術課題の整理

図 8（前出）では従来型の開発プロセスで進めた際の課題を列挙した。表 1（前出）の知識体系も開発プロセスに近い視点でまとめられている。それに対して、技術的な課題を抽出・整理するには、むしろシステム構成に着目した方が検討しやすい。

そこで、図 23 では、機械学習応用システムのおおまかな構成に対応させる形で、図 8 等であげられた課題を整理し直した。ここでは、機械学習型と従来型が混在した形でシステムの構成を考えている。このような整理をしてみると、技術課題を考える上で、次の 2 点が重要なことがわかる。

(i) 機械学習自体の品質はデータ集合が左右
(ii) 機械学習は 100% 保証ができないので、全体システムとしての対策が重要

また、AI ソフトウェア工学の問題に対して、産業界の問題意識･危機感が強く、産業界主導の取り組みは積極的に進められるものと考えられる。しかし、機械学習は本質的に確率的であり、解消できない不確かさを有するため、品質保証は根本的に難しい。従来手法の単純な改良での解決は困難であり、パラダイム転換に対応する新しい理論・原理を生み出すような学術研究が重要である。

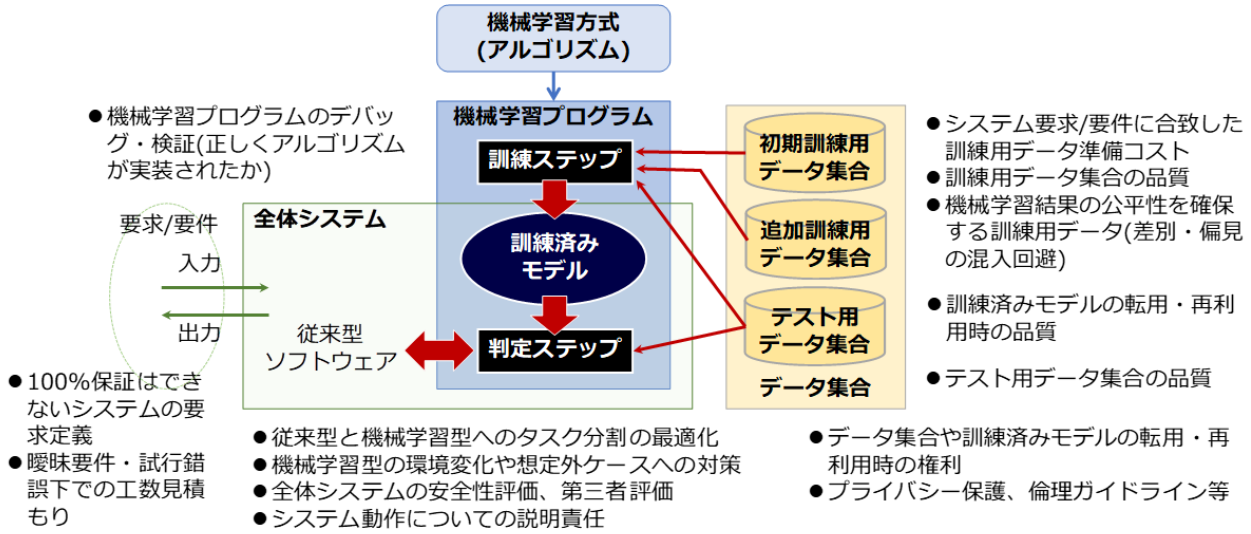


図 23　機械学習応用システムの構成と技術課題

このことから､AIソフトウェア工学は次のような2面からの総合的な取り組みが望まれる。
（A）実践から得られる知見の体系化のアプローチ
（B）理論・原理に関わる学術研究のアプローチ

以上のような検討に基づき、図23に示したような技術課題を（i）（ii）と（A）（B）という2つの軸で整理し直したものを表2に示す。すなわち、表2では、技術課題を（i）機械学習自体と（ii）全体システムという2タイプに分け、さらに、それぞれの技術課題に対して、（A）実践から得られる知見の体系化のアプローチと（B）理論・原理に関わる学術研究のアプローチという2つの面から考えられる取り組み方をあげた。

表２　技術課題に対する２面からのアプローチ

技術課題		(A) 実践から得られる知見の体系化のアプローチ	(B) 理論・原理に関わる学術研究のアプローチ
(i)機械学習自体	システム要求 / 要件に合致した学習用データ準備コスト	データ収集・加工・アノテーションの効率化ツールの整備	少量データや少量ラベル付きデータからの機械学習法の体系的構築
	訓練用データ集合の品質	訓練用データの量・分布特性に関するノウハウ集積、データ投入順の経験的最適化	訓練用データの品質評価 (想定外・脆弱性の最小化) の理論・指標
	テスト用データ集合の品質、機械学習プログラムのデバッグ・検証	テスト用データの量・分布特性に関するノウハウ集積	テスト用データの品質評価 (想定外・脆弱性の最小化) の理論・指標
	機械学習結果の公平性を確保する訓練データ (差別・偏見の混入回避)	学習時にセンシティブな説明変数の除外	公平性を確保するための機械学習の理論 (併せて解釈性も)
	訓練済みモデルの転用・再利用時の品質	訓練データ履歴の蓄積・管理	モデル適用環境の差分分析に基づくモデルや追加学習の品質評価
	訓練済みモデルの転用・再利用時の権利	訓練済みモデルの知財権・利用ガイドラインの整備	転用・再利用に関わる不正な利用の検知や品質の評価
(ii)全体システム	100% 保証はできないシステムの要求定義	要求定義プロセス (顧客の理解を得る) のノウハウ集積	リスク要因分析に基づく安全設計手法
	曖昧要件・試行錯誤下での工数見積もり	事例・経験の蓄積に基づく工数見積もり	不確かな要因の特定・解消を含む工数推定モデル
	従来型と機械学習型へのタスク分割の最適化	経験の集積による使い分け・デザインパターン化	機械学習型と記号推論型の最適統合原理の探求
	機械学習型の環境変化や想定外ケースへの対策	機械学習型の使用条件の経験的な限定、経験的なリカバリー処理	環境変化・想定外ケースの理論的判定とリカバリー処理の最適設計手法
	全体システムの安全性評価、第三者評価	従来型の安全性評価を部分修正して適用 (それが可能な範囲で限定運用)	機械学習型と従来型の混在システムの安全性評価の理論・指標
	システム動作についての説明責任	説明の必要なパートとブラックボックスが許容されるパートに分けて技術を使い分け	機械学習の解釈性向上、人間による受容性に関する理論

ここで、（A）については産業界主体で取り組みが進むに違いないが、（B）については基礎研究としての取り組みを強化する施策が望まれる。

そこで、表2の整理をもとに、（B）の内容を精査し、（B）の取り組みが必要あるいは特に有効なものを選び、技術的にある程度まとまった塊を抽出した。その結果を示したのが３－２節に示した図13である。基礎研究として重要な技術チャレンジとして、①機械学習自体の品質保証、②全体システムとしての安全性確保、③ブラックボックス問題への対策、④問題を効率的に解く工学的な枠組みという4つを抽出した。また、基礎研究課題の他に、国として検討が求められる課題として、安全性基準の策定（および第三者評価の仕組み）、訓練済みモデルの再利用に関わる知財権等も重要である。

表３　丸山・鵜林が示した機械学習工学における技術課題

丸山による［2］	（1）訓練データの準備と管理	データクレンジング、教師信号の付与（学習タイプで異なる）
	（2）生成物の再利用	訓練データセットの再利用､訓練済みモデルの再利用（Copy, Fine tuning, Ensemble, Distillation）､訓練済みモデル交換フォーマット
	（3）品質の担保	バグと精度の交互作用（CACE）、テスト、セキュリティ、訓練時からのデータ分布の変化、プロダクト品質とプロセス品質
	（4）解釈可能性 Interpretability	透明性（シミュレーション可能性、分解可能性、アルゴリズム透明性）、事後解釈可能性（テキストによる説明、可視化、局所データ影響による説明、例による説明）
	（5）要求工学	要求発生のタイミング、要求の厳密性
	（6）計算資源とその管理	CPU と GPU への振り分け、訓練時・推論時それぞれのワークロードへの最適化
	（7）ソフトウェアアーキテクチャ―パターン	演繹的開発 vs 帰納的開発､演繹的開発と帰納的開発の組み合わせ
鵜林による［29］	（1）機械学習のための要求工学	要求時のオープン性・曖昧性、顧客の期待値のコントロール
	（2）機械学習のための Reasoning	理由説明・解釈可能性
	（3）機械学習のためのソフトウェアアーキテクチャ―設計	機械学習モジュールは確率的であることを踏まえたシステム全体の品質保証
	（4）機械学習のためのテスト技術	訓練データセットの品質、テストの網羅性とは
	（5）機械学習システムにおけるモデルマネジメント	モデルのバージョン管理、トレーサビリティ
	（6）機械学習システムを継続的に運用する技術	初期学習時からの入力データの傾向変化、継続的な訓練済みモデルの改善
	（7）機械学習システム開発のためのプロセス、プロジェクト管理	工数見積もりの難しさ、多様なスキル人材のチームマネジメント
	（8）機械学習システム運用のための社会制度	100% の品質保証はできない、社会制度・社会契約、第三者評価
	（+1）MLSEBOK の作成	ソフトウェア工学の SWEBOK を機械学習の特性を考慮して拡張

なお、（A）（B）の区別はしていないが、AI ソフトウェア工学（機械学習工学）の技術課題は、他に論文［2］［29］でも示されている。参考のため、表 3 に示す。

（4）インタビューやワークショップで得た主なコメント・示唆

プロポーザルの提言骨子と重なるものは、ここでは省略する。プロポーザル本文で詳しく触れていないものを中心に、主なものを列挙する。

- 機械学習では、解消できない不確かさという問題が根本にあるのが、従来のソフトウェア工学との違い。従来ソフトウェアは、頑張れば完全にモデリングできる（不確かさは解消できる）と考えていた。
- 機械学習型のケースでは事前に条件分けを尽くすことができない。事前に想定したケースだけでなく､想定外が発生する。多面的に見張る､多重化して見張る等が必要。
- 命に関わる部分と利便性に関わる部分を分離。機械学習型のモジュールは 100% 保証ができないので､同じ機能に関する別アプローチや問題発生時のリカバリー策も併用する。
- 結果だけでなく入力データの分布も見張る。訓練時の環境から変化していると､訓練済みモデルが有効でなくなる。
- テスティングでは、正解から少しずらす、ノイズをかけたテストデータを作ることがポイント。

- Google 等、機械学習型システム開発を実用展開している部隊は品質問題に対しても何らかの取り組みをしているのは間違いない。ただし、アプリケーション依存で、学問的体系化はまだ。これに取り組むことが重要。
- あと 2 年もすれば、新しく作られる情報システムの少なくとも半数は、何らかの形で機械学習を用いた技術が含まれるようになると思う。
- 機械学習応用システムのライフサイクルをとらえ、データや訓練済みモデルの再利用にも着目する必要がある。
- AI の社会実装・事業化においてリスク最小化のマネージメントが重要。そのため機械学習型システムの品質管理基準（通常有すべき安全性とは？）が必要。
- 工業的規格がないと取引が困難。訓練プロセスを受託する事業を想定した場合、訓練済みモデルに対する品質基準（受入基準）をどうするか？　現状は曖昧なまま取引していることが多い。
- 検証が全部は不可能というのは、実は AI システムだけでなく従来システムでも同じ。デザインレビューが非常に重視されている。
- 従来もハードの故障やソフトのバグがあり、100% 保証はできないが、リスク受容の考え方ができていた。
- 100% 保証できない技術であることを含めて、要求定義について顧客の理解を得ることが必要。
- 費用積み上げ型の SI ビジネスから、顧客の受け取り価値ベースの成功報酬ビジネスへ変えていく。
- 非機械学習型との違いとして、その置き換え開発か / 新規開発かがアセスメントの困難さに、人のアクションを支援するのか / システムが自動実行するのかがデプロイメントの困難さにつながっている。
- 精度・信頼性を高める技術開発だけでなく、データを集めてアノテーションするという泥臭い作業の方法論も重要。
- 使われながら機械学習して動作が変わるのは怖いので（Tay の事件があったので）、それを許可する顧客はいない。事前に用意されたパターンのランキングを変える程度。
- システム全体がデータによる帰納的な動作定義になることはなく、ソフトウェア（コード）による動作定義がメインで、データによる帰納的な定義は部分的に使われる。それらの組み合わせとトップダウンなデザインの考え方が重要。
- 安全性・信頼性は、ソフトウェアでなくシステムに対して要求されるものであり、システム仕様によって要求レベルも異なる。
- 自動運転分野の検討が先行しているが、競争が激しく多くが非公開のため、規格を定めるのが難しい。
- 機械学習応用システムの品質評価に関して、プロダクト品質、サービス品質、プラットフォーム品質を区別して論じる必要がある[31]。そのため、ターゲット応用もいくつかの軸で分類して、要求される品質レベルを論じるべき。
- ソフトウェアの保守には是正保守と適応保守がある。プラットフォーム品質は適用保守に該当する。機械学習では環境変化に適応して進化していくことが必要で、産業界が本当に欲しいのはプラットフォーム品質ではないか。
- 品質保証、ソフトウェア開発プロセス、機能安全という 3 つが別々に発展してきたが、

これら本質的に同じなので融合し、また、ハードウェア、ソフトウェア、AI も統合して考えることが必要。

- ハードウェアの故障は確率的で、ソフトウェアのバグは決定論的と言われる。しかし、ソフトウェアについても、実際はテスト工数の限界によって確率的な面が入るので、AI に限らず確率的な扱いは必要。
- ソフトウェア工学はモデルの塊で、そのモデル作りは人力･労働集約的な仕事。自分たちの仕事の情報化が進められていない。デジタルトランスフォーメーションが遅れている。
- 安全性･信頼性を（a）プロセス的（正しく作られているか）、（2）プロダクト的（挙動が安全か）、（3）説明的（結果を説明できるか）の 3 面から考える。
- ソフトウェア工学分野の研究者と産業界が一体となって研究を進めるべき。米国の数学分野が応用数学を重視し、現実問題を取り入れたことで発展した。ソフトウェア工学でも同じことが言える。

付録２．国内外の状況

（１）産業界

問題意識が急速に高まっているが、根本的な解決策はまだ見えていない状況である。

そのような状況において、自動運転分野での検討が先行している。欧州の PEGASUS プロジェクト、デンソーの AI 品質基盤、宇宙航空研究開発機（JAXA）と日本自動車工業会（JAMA）の共同検討等が進んでいる。

（２）AI 研究コミュニティー

機械学習の精度・性能を高める競争が激しく、総じて開発法への関心は低い。

ただし、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･限界テスト（Adversarial Examples 等）、ブラックボックス問題が検討され始めた。米国 DARPA が XAI（説明可能な AI）への研究開発投資プログラムを 2017 年から開始した。2014 年から毎年開催されている FAT/ML（Fairness, Accountability, and Transparency in Machine Learning）ワークショップ等で、機械学習における公平性・説明責任・透明性等が議論されるようになった。また、深層学習と記号推論の融合が次の重要課題として議論されている（2018 年 1 月に東京大学で開催された「深層学習の先にあるもの－記号推論との融合を目指して」公開シンポジウム等）。

これらは開発法そのものの研究ではないが、開発法を設計する上での指針になり得る。

（３）ソフトウェア工学研究コミュニティー

国内では、3 ～ 4 年前から兆し[5]が見られたが、2017 年初頭から学会・業界イベントでの基調講演や企画セッション等が立て続けに開催され（例えば以下の通り）、一気にホットトピック化した。

- 2017 年 2 月 3 日：情報処理学会ソフトウェアジャパン 2017
  招待講演「演繹から帰納へ：情報技術のパラダイムシフト」（丸山宏）
- 2017 年 8 月 31 日：情報処理学会ソフトウェア工学シンポジウム SES2017
  基調講演「設計から学習へ：かわりゆく人工知能ソフトウェア」（山川宏）
  パネル討論「機械学習とソフトウェア工学」（青山幹雄、丸山宏、濱田晃一、石川冬樹）
- 2017 年 10 月 4 日：CEATEC JAPAN 2017 産総研人工知能コンファレンス
  招待講演「ソフトウェア工学からみた機械学習」（中島震）
- 2017 年 12 月 14 日：JEITA ソフトウェアエンジニアリング技術ワークショップ 2017
  ～人工知能ブームの中でのソフトウェアエンジニアリング～
  基調講演「機械学習工学に向けて」（丸山宏）
- 2018 年 1 月 18 日・19 日：情報処理学会ソフトウェア工学研究会ウィンターワークショップ 2018 in 宮島

5　例えば、2015 年 11 月 24 日に開催された日本学術会議公開シンポジウム「by 機械学習、of 機械学習」における岡野原大輔（Preferred Networks）の講演「機械学習による機械の学習：自動運転からロボットの協調生産まで」の中で、機械学習がシステム開発のパラダイム転換を起こすことが言及された。また、ほぼ同時期、丸山宏（当時 統計数理研究所、現在 Preferred Networks）や中島震（国立情報学研究所）らからも、機械学習をソフトウェア工学から見た技術課題等が指摘されていた。

「機械学習工学」セッション

- 2018 年 6 月 8 日：人工知能学会全国大会 JSAI2018
  企画セッション「機械学習工学とは－機械学習システムを創り上げるための工学的課題 －」
- 2018 年 8 月 29 日～ 31 日：日本ソフトウェア科学会第 35 回大会
  「機械学習工学」セッション
- 2018 年 10 月 18 日：CEATEC JAPAN 2018 産総研人工知能コンファレンス
  招待講演「機械学習ソフトウェアの品質：プロダクト，サービス，プラットフォーム」（中島震）
- 2018 年 12 月 12 日：情報処理学会連続セミナー 2018 第 6 回
  「人工知能時代のソフトウェアエンジニアリング」（本橋洋介、丸山宏、高柳慎一、門田暁人、中島震、福島俊一）

さらに、2018 年 4 月には、日本ソフトウェア科学会に機械学習工学研究会 MLSE が発足し、5 月 17 日に開催された MLSE キックオフシンポジウムには 500 名を超える参加者が集まった。参加申し込みが短期間で満席になり、産業界からの参加者数が多く、問題意識の高さが見られた。さらに、7 月 1 日・2 日に泊りがけで開催された第 1 回 MLSE ワークショップ（MLSE2018）も満席となり、活発な議論が交わされた。

同じ 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。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う。

また、宇宙航空研究開発機構（JAXA）と日本自動車工業会（JAMA）は、AI 時代のソフト品質保証の考え方に関する連携を進めると、2018 年 3 月 9 日にプレスリリースを行った。安全性評価解析技術（STAMP/STPA[19]等、制御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解析手法）や IV&V（Independent Verification and Validation：システムのリスク低減を目的として開発者と異なる視点からリスクを抽出し対処する活動）[30]を活用し、AI による自律化システムの安全のためのソフトウェア品質確保の考え方をまとめるということである。

このように、日本は、開発法という視点からいち早く検討を開始し、具体的な活動において、海外にやや先行している。

一方、海外においては、2017 年後半から、同様に、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、新しいパラダイムは「Software 2.0」（Tesla による）[3]や「Machine Teaching」（Microsoft による）[4]と呼ばれ、国際学会の基調講演等で「Software 2.0」が取り上げられるようになった。

- 2018 年 6 月：ISCA2018 にて Kunle Olukotun（スタンフォード大学）が「Designing Computer Systems for Software 2.0」と題した基調講演
- 2018 年 6 月：Deep Learning World Las Vegas 2018 にて Gil Arditi（Lyft）が「Deep Learning as Software 2.0」と題した講演

また、AI ソフトウェア工学に関する問題意識が議論される機会も増えており、ワークショップや関連組織の動きも活発化している。

- カナダでは、モントリオール工科大学に SEMLA イニシアティブ（The Software Engineering for Machine Learning Applications initiative）が発足し、2018 年 6 月 12 日・13 日にキックオフシンポジウムを開催
- 2018 年 7 月 15 日に Joint Workshop between ICML, AAMAS and IJCAI on Deep（or Machine）Learning for Safety-Critical Applications in Engineering が開催
- 中国では、2018 年 11 月 23 日～ 25 日に深圳で開催される第 17 回全国ソフトウェア応用大会において、「AI に基づくソフトウェア工学」のセッションが設けられる

（4）AI 品質関連の標準化活動や安全基準策定の動向

国際標準化機構 ISO と国際電気標準会議 IEC 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いる（JTC：Joint Technical Committee、SC：Subcommittee）。これらの SC で 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 SC7 下の WG6 がソフトウェア品質、WG26 がソフトウェア・テスティングを扱っており、さらに自律システムの倫理等を扱う WG も新設される予定である。

また、ISO TC159 で人間工学、その下の SC4 でヒューマンコンピューターインタラクションを扱っており、AI を使う際の利用者視点でのガイドラインを作ろうとしている。

IEEE Standard Association でも Global Initiative for Ethical Considerations in AI and Autonomic Systems が議論されている。IEEE P7000 シリーズがその規格になる。

他にも、米国ベンチャーの Edge Case Research と、米国の認証機関である Underwriters Laboratories が共同で自律システムの安全性基準を策定しようとしている。

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。

Google DeepMind は自社の AI 開発ガイドライン[34]を、仕様、頑健性、保証という 3 面からまとめていることを発表した。

国内では、ソフトウェア品質知識体系 SQuBOK（Software Quality Body of Knowledge）策定部会が、2020 年に発行予定の次期バージョン V3 において、AI 応用システムの品質も取り上げる予定である[35][36]。

付録３．専門用語説明

● 機械学習（Machine Learning）

人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピューターで実現しようとする技術・手法のことである。コラム②を参照。

● 記号推論

機械学習がデータ駆動型でボトムアップ的アプローチであるが、記号推論はトップダウンから論理や知識を利用して記号を用いて推論しようとするもの。両者を融合することで、それぞれの強みと弱みをカバーできる可能性がある。

● コンテキストアウェアシステム（Context-Aware Systems）

過去の行動履歴、現在の時刻・スケジュール・位置情報に基づいて、次の行動に必要な情報を提供するシステムのことをいう。

● 深層学習（Deep Learning）

多層のニューラルネットワーク（ディープニューラルネットワーク）による機械学習手法。深層学習登場以前、４層以上の深層ニューラルネットは、技術的な問題によって十分訓練させられず、性能も芳しくなかった。しかし、近年、ヒントンらによる多層ニューラルネットワークでの機械学習の研究や、訓練に必要な計算機の能力向上、および、Web の発達による訓練データ調達の容易化によって、十分訓練させられるようになった。その結果、音声・画像・自然言語を対象とする問題に対し、他の手法を圧倒する高い性能を示し、2010 年代に普及した。

● FRAM（Functional Resonance Analysis Method）

2004 年に南デンマーク大の E. Hollnagel が開発した機能共鳴の分析による安全解析手法の一つである。システムの機能間の依存性や結合性の様子をモデル化し、分析する。各機能を６つの要素で定義し、各機能間が互いにどの要素とインタラクションするのかを明らかにし、その関係性の中に安全に係わるシステムの長所や短所を見出す。

● ILSVRC（The ImageNet Large Scale Visual Recognition Challenge）

ILSVRC は 2010 年から始まった大規模画像認識の競技会である。開始当初は、「画像分類」だけだったが、現在では、「画像内で物体がどこにあるのかを推測する」、「画像内で指定された 200 カテゴリーの物体を検出する」、「動画内で指定された 20 カテゴリーの物体を検出する」の３つのコンテストが開催された。

● STAMP/STPA

STAMP（System Theoretic Accident Model and Processes）/STPA（System Theoretic Process. Analysis）は、マサチューセッツ工科大学の Nancy Leveson 教授が提唱した、大規模・複雑化するシステムに適した新しい安全解析手法。従来のシステム構成要素のすべ

ての故障を明示化し、低減するという手法だけでは限界があるため、事故を回避するために制御行動の乱れを分析し、それを防ぐ手法と組み合わせることによって安全性向上を図ろうとする手法。

● SWEBOK（Software Engineering Body of Knowledge）

SWEBOK とは、米国電気電子技術者協会（IEEE）および米国計算機学会（ACM）によってまとめられた、ソフトウェア工学（ソフトウェアエンジニアリング）に関する知識の分類体系である。SWEBOK における知識体系は、「SWEBOK ガイド」としてまとめられている。ソフトウェアエンジニアリングの知識の全体像を把握し、効率的に学習するための手引きして利用されている。

● V 字モデル（V-Model）

ドイツ政府関連のソフトウェア開発工程を規定するために開発された。V 字型に表される概念図の左側はシステムの仕様を記述していく流れを示し、右側はテストの流れを示す。それぞれの同じ高さの部分は開発の詳細さのレベルを表している。V 字の左側に対する右側は、上から、ユーザー要求仕様⇔受け入れテスト、機能仕様⇔操作テスト、設計仕様（詳細仕様）⇔性能テスト、となる。

● CACE（Changing Anything Changes Everything）

機械学習は、複数のデータ源を混ぜ合わせるため、絡み合いを生じ独立した改良を不可能にする。たとえばデータやパラメータが少しでも変われば、予測結果が少し変わる、あるいはは劇的に変わる、ということがおこる。

● Adversarial Examples

深層学習の不可解な挙動のことで、adversarial attack や、adversarial perturbation 、画像関係の場合は adversarial image と呼ばれることもある。分類器に対する脆弱性攻撃のようなもの。例えば、分類器が正しく分類できていた画像に、人の目では判別できない程度のノイズをのせることで、作為的に分類器の判断を誤らせることができる。

付録４．参考文献

［1］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。http://jssst.or.jp/files/user/taikai/2017/GENERAL/general6-1.pdf
［2］丸山宏・城戸隆：「機械学習工学へのいざない」、人工知能学会誌、33 巻 2 号、2018 年 3 月。
［3］Andrej Karpathy: “Software 2.0”, Medium, 2017.11.12. https://medium.com/@karpathy/software-2-0-a64152b37c35
［4］Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang and John Wernsing: “Machine Teaching: A New Paradigm for Building Machine Learning Systems”, arXiv:1707.06742, 2017. https://arxiv.org/abs/1707.06742
［5］福島俊一・藤巻遼平・岡野原大輔・杉山将：「ビッグデータ×機械学習の展望：最先端の技術的チャレンジと広がる応用」、情報管理、60 巻 8 号、2017 年 11 月。https://doi.org/10.1241/johokanri.60.543
［6］D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary and Michael Young: Machine Learning: “The High-Interest Credit Card of Technical Debt”, Google Publications, 2014. https://ai.google/research/pubs/pub43146
［7］福島俊一：「機械学習工学における研究課題の概観」、第 1 回機械学習工学ワークショップ MLSE2018、2018 年 7 月。https://drive.google.com/file/d/1Yp48Uf9RX64b2--1M_zInvpVK_9nmi3i/view
［8］Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno and Dawn Song: “Robust Physical-World Attacks on Deep Learning Models”, arXiv:1707.08945, 2017. https://arxiv.org/abs/1707.08945
［9］「マイクロソフトの機械学習 AI「Tay」、ネットで差別と陰謀論に染まって一日で公開停止」、engadget、2016 年 3 月 25 日。https://japanese.engadget.com/2016/03/24/ai-tay/
［10］平和博：「AI と「バイアス」：顔認識に高まる批判」、HUFFPOST、2018 年 9 月 3 日。https://www.huffingtonpost.jp/kazuhiro-taira/ai-20180903_a_23514984/
［11］Jeffrey Dastin：「焦点：アマゾンが AI 採用打ち切り、「女性差別」の欠陥露呈で」、REUTERS、2018 年 10 月 11 日。https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN
［12］廣瀬明倫:「人工知能が経営にもたらす「創造」と「破壊」」、EY 総合研究所、2015 年 9 月。https://www.shinnihon.or.jp/shinnihon-library/publications/issue/eyi/knowledge/fsi/pdf/2015-09-15.pdf
［13］青山幹雄:「機械学習工学への期待～機械学習が工学となるために～」、日本ソフトウェア科学会 機械学習工学研究会 キックオフシンポジウム、2018 年 5 月 17 日。https://www.slideshare.net/MLSE/ss-97568695?ref=https://mlxse.connpass.com/event/80434/presentation/

［14］石川冬樹:「AI 時代における品質保証のチャレンジ～機械学習の難しさと（AI による）テスティング」、ESTIC2018（Embedded System Technology and Innovation Conference）、2018 年 7 月 20 日。
http://research.nii.ac.jp/~f-ishikawa/work/1807-ESTIC18-AI+Testing.pdf
［15］中島震 :「データセット多様性のソフトウェア・テスティング」、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。
http://jssst.or.jp/files/user/taikai/2017/FOSE/fose3-2.pdf
［16］明神智之 :「AI 搭載システムの品質保証」、ソフトウェアテストシンポジウム JaSST'18 Tokyo、2018 年 3 月 8 日。
http://jasst.jp/symposium/jasst18tokyo/pdf/C5-1.pdf
［17］神嶌敏弘 :「Fairness-Aware Data Mining」、2018 年 3 月。
http://www.kamishima.net/archive/fadm.pdf
［18］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
［19］「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（IPA）、2016 年 3 月。https://www.ipa.go.jp/files/000051829.pdf
［20］「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（IPA）、2016 年 3 月。https://www.ipa.go.jp/files/000065199.pdf
［21］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」、NIKKEI Robotics、2018 年 6 月号。
［22］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 MLSE2018、2018 年 7 月。
https://drive.google.com/file/d/1RrIv3B3hOJQM8YawLD5E44c1BeHzh5GA/view
［23］Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi and Fosca Giannotti: “A Survey Of Methods For Explaining Black Box Models”, arXiv:1802.01933, 2018. https://arxiv.org/abs/1802.01933
［24］松下外:「GDPR における人工知能（AI）の取扱いについて（2）」、イノベンティア、2018 年 8 月 6 日。https://innoventier.com/archives/2018/08/6753
［25］David Gunning: “Explainable Artificial Intelligence（XAI）”, DARPA Program Update, November 2017. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
［26］「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」、NIKKEI ELECTRONICS、2018 年 9 月号。
［27］David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel and Demis Hassabis: “Mastering the game of Go with deep neural networks and tree search”, Nature, Vol. 529, 2016. https://www.nature.com/articles/nature16961
［28］我妻広明 :「データ駆動型人工知能と論理知識型人工知能の融合による解釈可能な自動運転システムに関する研究」、NEDO「次世代人工知能・ロボット中核技術開発」（人工知

能分野）中間成果発表会、2017 年 3 月 29 日。
https://www.airc.aist.go.jp/info_details/docs/170329/1730-Wagatsuma.pdf
［29］鵜林尚靖：「機械学習工学が解決すべき 8 つの研究課題」、第 1 回機械学習工学ワークショップ MLSE2018、2018 年 7 月。
https://drive.google.com/file/d/1j8xMp3HDBEWhqbhoiV6WqQLLEilePyVG/view
［30］「JAXA IV&V の概要 ～ IV&V と評価戦略可視化の関係～」、宇宙航空研究開発機構、2016 年 1 月 19 日。https://www.ipa.go.jp/files/000050256.pdf
［31］中島震：「機械学習ソフトウェアの品質：製品、サービス、プラットフォーム」、電子情報通信学会 知能ソフトウェア工学研究会（KBSE）、2018 年 11 月 9 日。
［32］Sandeep Neema: “Assured Autonomy”, DARPA Program, August 2017.
https://www.darpa.mil/attachments/AssuredAutonomyProposersDay_Program%20Brief.pdf
［33］Keiichi Kisamori, Takashi Washio, Yoshio Kameda and Ryohei Fujimaki: “A Rare and Critical Condition Search Technique and its Application to Telescope Stray Light Analysis”, Proceedings of the 2018 SIAM International Conference on Data Mining（SDM18）, May 2018. https://doi.org/10.1137/1.9781611975321.64
［34］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
［35］鷲崎弘宜：「アジャイル品質保証の知識体系 − SQuBOK 2020 予定より −」、enPiT-Pro スマートエスイーセミナー、2018 年 9 月 7 日。
https://www.slideshare.net/hironoriwashizaki/squbok-2020
［36］SQuBOK 策定部会編：「SQuBOK Review 2018」、2018 年。https://www.juse.or.jp/sqip/squbok/file/squbok_rev2018_20181016.pdf

■戦略プロポーザル作成メンバー■

総括責任者	木村 康則	上席フェロー	（システム・情報科学技術ユニット）
チームリーダー	福島 俊一	フェロー	（システム・情報科学技術ユニット）
チームメンバー	相田 俊一	主任調査員	（未来創造研究開発推進部 推進第３グループ）
	奥山 隼人	主査	（未来創造研究開発推進部 推進第１グループ）
	川村 隆浩	主任調査員	（情報企画部 システム高度化グループ）
	國岡 崇生	課長代理	（人財部 職員課）
	高地 伸夫	主任調査員	（戦略研究推進部 ICT グループ）
	澤田 朋子	フェロー	（海外動向ユニット）
	嶋田 義皓	フェロー	（システム・情報科学技術ユニット）
	蒋 赫	主査	（イノベーション拠点推進部 共創グループ）
	坂内 悟	フェロー	（システム・情報科学技術ユニット）

※お問い合せ等は下記ユニットまでお願いします。

CRDS-FY2018-SP-03

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

STRATEGIC PROPOSAL

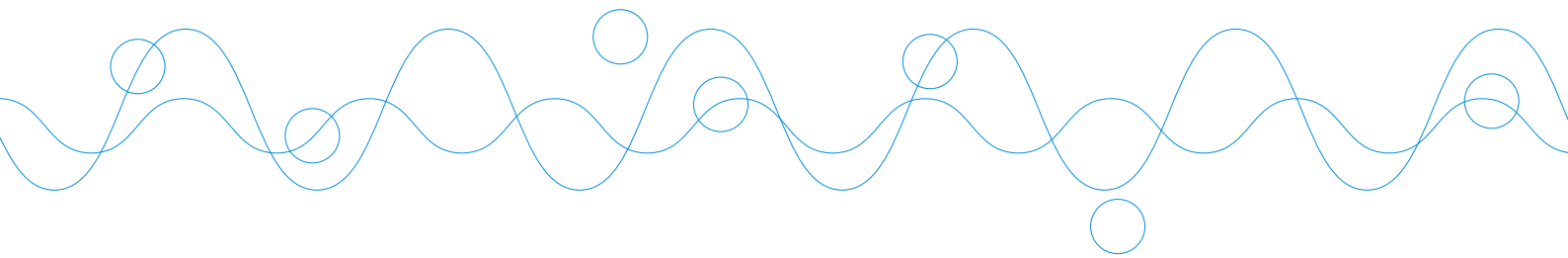
Toward new-generation software engineering to guarantee safety and reliability of AI application systems

平成 30 年 12 月　December 2018
ISBN 978-4-88890-615-9

国立研究開発法人科学技術振興機構研究開発戦略センター
システム・情報科学技術ユニット
Systems and Information Science and Technology Unit
Center for Research and Development Strategy,
Japan Science and Technology Agency

〒102-0076　東京都千代田区五番町 7 番地
電　　話　03-5214-7481（代表）
ファックス　03-5214-7385
http://www.jst.go.jp/crds/

©2018 JST/CRDS

許可無く複写／複製することを禁じます。
引用を行う際は、必ず出典を記述願います。
No part of this publication may be reproduced, copied, transmitted or translated without written permission. Application should be sent to crds@jst.go.jp. Any quotations must be appropriately acknowledged.

ISBN 978-4-88890-615-9