2210AIquality
2022.10　ブログ：『AI 品質保証にかかわる国内外の取り組み動向』を読んで、の詳細
（→ http://www.1968start.com/M/blog/index2.html#2210　）

『AI 品質保証にかかわる国内外の取り組み動向』を読んで

中所武司

■この本の読書のきっかけ

情報処理学会誌の最新号に掲載されていた下記の解説に興味を持った。
特集：AI の品質保証：「1．AI 品質保証にかかわる国内外の取り組み動向」
情報処理, 63(11)， e1-e6 (2022-10-15)

■記事内容の要約とコメント（→★）

【AI 応用システムの品質問題】

・AI 適用産業は，交通・医療のような人命にかかわる分野にも広がり，
その品質にかかわる AI 特有の問題が指摘されるようになってきた．
（代表例）
（a）**AI のブラックボックス問題**
深層学習などのニューラルネットワーク型の機械学習モデルは，構造が複雑で，
どのような条件のときにどのような動作をするか，明確な動作保証ができず，
事故などが起きたときの原因解明にも難しさが生じる．

→★ＡＩの説明機能については、以下のブログで言及：
2022.8　「ＡＩ判断の根拠を説明するＸＡＩ」を読んで
http://www.1968start.com/M/blog/index2.html#2208

（b）**AI のバイアス問題**
機械学習に用いる学習（訓練）データの偏りが不公平な判定結果を生むことがある．
人種や性別に応じて判定に不公平が生じた事例がしばしば問題視されている．

→★これは、ＡＩの技術的な学習能力の問題ではない。
人間社会でも、偏見は世代間で受け継がれてきた。

（c）**AI の脆弱性問題**
敵対的サンプルと呼ばれる手法：機械学習ベースの画像認識システムでは，
対象画像に，人間にはあまり差異を感じにくい程度の加工（ノイズ）を加えることで，
その画像の認識結果を変えることができる．

別の事例として，利用者との対話からオンライン機械学習を行うチャットボットが，悪意のある利用者との対話によって，不適切な発言を繰り返すようになってしまい，わずか１ 日で運用停止に追い込まれたこともよく知られている．

→★最後のチャットボットの例は、(c)の脆弱性ではなく、(b)のバイアス問題と思われる。
この例については、以下の過去ブログで言及している。

2016.3　**差別発言した人工知能の学習モデルは？**
http://www.1968start.com/M/blog/old.html#1603c
2017.8　**共産党批判でサービス停止のチャットボットの再教育とは？**
http://www.1968start.com/M/blog/index.html#1708
2018.6　**差別発言をしたチャットボットＴａｙ（２年前）への追加コメント**
http://www.1968start.com/M/blog/index.html#1806b

【システム開発のパラダイム転換】

・従来のシステム開発はプログラミングによる演繹型の開発法であるのに対して，
AI 応用システム開発は，コアな部分に機械学習による帰納型の開発法を用いるので、
システム全体としては帰納型と演繹型を組み合わせた開発になる．

・従来の開発では，明示的に書いた手続きや判定ルールで，システム動作が決まるが，
機械学習では，例示データをまねて自動的に判定ルールができ，システム動作が決まる．

・そこで，品質保証のために重要なシステムテストについて、

＊演繹型の開発法では，想定されるケースを列挙し，すべてのケースについて，
不適切な事態を引き起こさないかを，順にチェックしていけばよい．
＊一方，帰納型の開発法では，列挙したケースの動作仕様を事前に決められないので，
どれだけのケースをテストでカバーすれば十分かの判断が難しい．
＊また，学習（訓練）データの追加で，システム全体の動作が変わり得ることが，
テストやデバッグを難しいものにしている．

→★この問題については、次のブログでコメントしている：

2018.4　機械学習工学とソフトウェア工学の共通点と相違点
http://www.1968start.com/M/blog/1804MLvsSE.pdf
2019.1　ＡＩシステム検証へのニューロンカバレッジの有用性について
http://www.1968start.com/M/blog/index.html#1901

【AI 関連ガイドライン】

・AI の ELSI（倫理的・法的・社会的課題）が国家レベルで議論されるようになった．
2019 年に， OECD や G20 など国際連携の枠組みでの合意による AI 原則が宣言された．

・日本政府は，内閣府が「人間中心の AI 社会原則」を 2019 年 3 月に公表した．
この中では，人間中心の原則，教育・リテラシーの原則，プライバシー確保の原則，
セキュリティ確保の原則，公正競争確保の原則，公平性・説明責任・透明性の原則，
イノベーションの原則という 7 つが挙げられている．

・欧州委員会では、2019 年 4 月に「信頼できる AI のための倫理指針」を公表した．
米国では，2017 年 1 月にアシロマでの会議で公表した「アシロマ AI 原則」や，
IEEE が 2019 年 3 月に公表した「倫理的に配慮されたデザイン」がある．
中国政府も、2019 年 6 月には「次世代 AI ガバナンス原則」を公表した．

・これらは抽象度の高い原則レベルのもので，あまり大きな差は見られない．
そして，取り組みのフェーズは，原則から実践へと移行した．
実践フェーズでの注目は，欧州委員会の 2021 年 4 月の「AI 規制法案」である．
AI のリスクを，(a) 容認できないリスク，(b) ハイリスク，(c) 限定的なリスク，
(d) 最小限のリスク／リスクなし，の 4 段階に分類し，(a) の AI は使用禁止，
(b) は事前に適合性評価，(c) は透明性の確保が必要とした．

・日本は，2019 年 5 月に AI プロダクト品質保証（QA4AI）コンソーシアムから
「AI プロダクト品質保証ガイドライン」，2020 年 6 月に産業技術総合研究所から
「機械学習品質マネジメントガイドライン」が公開され，拡充改訂も行われている．

【研究開発動向】

・前述したパラダイム転換に対応した新たな技術開発が必要になり，
ソフトウェア工学と AI とが融合した新しい研究開発分野が生まれることになった．

・日本では，その研究コミュニティとして，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，QA4AI コンソーシアムも発足した．

・産業界では，AI 応用システム開発・運用全般にわたって効率化・最適化する考えで，
従来の DevOps フレームワーク（開発側 Dev と運用側 Ops の協調した取り組み）を
機械学習（ML）の開発・運用に発展させたものを MLOps と呼ぶようになった．

・また，さまざまな AI 応用システム開発を通して，繰り返し利用可能な問題解決策や

ノウハウをまとめた「機械学習デザインパターン」が共有・活用されるようになった.

・以下，冒頭で例示した 3 種類の問題への対策を中心に，技術をいくつか紹介する.

（a）**AI のブラックボックス問題**に対して，
　説明可能 AI（XAI）の技術では、深層学習のようなブラックボックス型モデルを，
決定木のような解釈性の高いモデルで近似する手法や，個々の判定結果について，
重視する特徴や関連深い学習データを特定し，判断根拠を近似する手法が開発されている.
対象に関する物理モデルや分野知識を事前に与え，それを制約とした学習を行うことで，
解釈しやすい結果を得る手法も開発されている.

→★（再掲）ＡＩの説明機能については、以下のブログで言及：
　2022.8　「ＡＩ判断の根拠を説明するＸＡＩ」を読んで
　　http://www.1968start.com/M/blog/index2.html#2208

（b）**AI のバイアス問題**に対して，
　公平性配慮型機械学習の技術では，公平性の指標（グループ公平性・個人公平性）を
定義し，各応用における公平性の状況を可視化する．このとき，複数の公平性指標は
同時に満たすことができず，かつ，公平性と精度はトレードオフ関係にあることから，
応用ごとに適切な目標設定が必要になる.

→★「複数の公平性指標は同時に満たすことができず」の具体例は？
　冒頭の事例の人種と性別の公平性は同時に満たせるのでは？

（c）**AI の脆弱性問題**を中心に，ブラックボックス問題やバイアス問題の関連技術として，
　機械学習のテスティングの技術開発では，起こり得るケースを網羅することが
原理的にできず，テストの正解を与えるシステム仕様も存在しないという前提のもと，
テストすべきケースを効率良くカバーしたテストを繰り返していくことが必要になる.
　ニューラルネットワーク内の活性化範囲を調べ，それを広げるようにテストパターンを
生成するニューロンカバレッジ法，入力を変えると出力が変わるという関係に基づき，
既存のテストケースから摂動によって多数のテストケースを生成するメタモルフィック
テスティング法，ヒューリスティックな探索を用いて，欲しいテストケースを表す
スコアを最大化するようなテストケースを生成するサーチベースドテスティング法など，
さまざまなテスティング技術の研究開発が進められている.

→★「起こり得るケースを網羅することが原理的にできず」は、従来型も同じでは？
　「システム仕様も存在しない」が、学習データをシステム仕様と考えれば、
　学習データをテストデータとして用いて、全問正解ならば、開発完了では？

もし、新たなデータで意図した出力結果がでなければ、システム仕様としての
学習データが不完全だったので、このデータを学習データに加えて再開発する。

→★ニューロンカバレッジ法については、次のブログでコメントしているが、
ニューラルネットワーク内の不活性化の部分を調べ，それを活性化するように
テストパターンを生成するのは難しいと思う：

2019.1　ＡＩシステム検証へのニューロンカバレッジの有用性について
http://www.1968start.com/M/blog/index.html#1901

【政策動向】

・各国政府は AI の品質保証・安全性確保も科学技術政策上の重要課題に位置付けている．
米国の DARPA の XAI プロジェクト（2017～2021）は，軍事における意思決定支援として
位置付けられている．さらに，Assured Autonomy プロジェクト（2018～2022）では，
自動運転車やドローンなどの自律システムの安全性確保が検討されている．

・日本政府は，2019 年 6 月に「AI 戦略 2019」を決定し、中核的課題として
「信頼される高品質な AI」がある．国の戦略投資による研究開発プログラムとしても，
科学技術振興機構や新エネルギー・産業技術総合開発機構による「信頼される AI」
「AI 信頼性」などのプログラムが推進されている．

【業界動向】

・AI 品質・AI 倫理にかかわる国際標準化活動も進められている．
IEEE 標準化協会は，「IEEEP7000 シリーズ」の規格策定を進めている．
ISO（国際標準化機構）と IEC（国際電気標準会議）は，AI の信頼性，ガバナンス，
テスティングを含む AI 関連標準化活動（ISO/IEC JTC 1/SC 42）を進めている．
米国 NIST は，2022 年 3 月に「AI RiskManagement Framework」を公開した．

・自動車業界の AI 応用では，誤認識や未学習ケースなど，故障以外の要因でリスクが
多々発生する．そこで，新たに ISO/PAS 21448（SOTIF）が策定され，2019 年 1 月に
公開された公開仕様書では、SOTIF（Safety of the Intended Functionality）は，
懸念されるケース・条件での動作が適切かを一つひとつ検討するアプローチをとる．
また，ドライバーが操作しない自動運転レベル 4 以上を想定した安全規格として，
米国の認証機関 Underwriters Laboratories は 2020 年 4 月に UL4600 を公開した．

【今後の展望・課題】

・AI の品質保証にかかわる背景と国内外の取り組み状況を概観した．
第 3 次 AI ブームで AI 応用システムの市場が拡大し，AI 品質保証問題について

産業界が危機感を持ち始めた 4 年前から研究開発と開発現場での実践が一気に進んだ．

・しかし，AI 品質保証の方法論・技術体系が確立できたとはまだ言えない．
確率的に振る舞う機械学習に対して 100% 保証を与えることは原理的に無理がある．
機械学習そのものとそれを含むシステム全体の両面から問題を丁寧に整理し，
問題が起きた後の対応も含めて実践的な対策を積み上げ，開発・運用プロセスについての
信頼・社会受容を獲得していくことが必要であろう．

→★「原理的に無理」については同感。しかし、ＡＩシステムを組み込んだ機器が
事故を起こせば、製造物責任法の対象になると思う。

・AI が自律性を持つことをどの程度まで受容するかについては議論があるところだが，
これは品質保証にも大きくかかわってくる．出荷後にオンライン機械学習を通して
モデルが変化する場合の品質保証や，擬人化 AI に対する人間の心理（過剰な能力期待）
と安全性・信頼性のかかわりなども，今後考えていく必要があろう．

→★従来のソフトウェア工学では、変更があれば、テストのやり直しが必要とされる

以上