機械学習品質マネジメントにおける AI の新潮流への対応
【機械学習品質マネジメントガイドライン 第 4 版 Annex】
要旨

産総研の機械学習品質マネジメント検討委員会は、2020 年に機械学習品質マネジメントガイドライン第 1 版を公表し、社会での安全・安心な AI 活用の実現に貢献してきた。本文書は大規模言語モデルなどの、基盤モデルに基づく生成系 AI の急速な発展を踏まえて、機械学習 AI システムの品質マネジメントの今後の課題と対応の見通しを示す。これに基づき、産総研は AI の新潮流の下でも継続して品質マネジメントへの貢献に取り組んでいく。

主な課題は以下の二つである。1）基盤モデルにより従来よりも複雑になったサプライチェーン上のステークホルダー間の役割分担。2）基盤モデルの訓練に用いたデータが基盤モデルの出力に原形に近い形で再現されることへの対処。両課題とも、外部から提供される基盤モデルを製品開発者が利用する場合、元の基盤モデルの生成に用いたデータの性質など、基盤モデル提供者しか知り得ない情報を品質管理に必要とすることから、ステークホルダー間で品質に関わる情報を必要な範囲で共有し、対処する必要がある。

このような観点からサービス開発者・提供者は、有償・無償を問わず可能な範囲で、自らが品質マネジメントを行うために必要十分な情報を提供している基盤モデル開発者により提供される基盤モデルを選択して用いることが望ましい。また、基盤モデル開発者が積極的にそのような情報を市場に提供することにより、社会全体での品質マネジメントの取り組みが進むことが望ましい。

機械学習品質マネジメントにおける AI の新潮流への対応

【機械学習品質マネジメントガイドライン 第 4 版 Annex】

2023 年 12 月 12 日

機械学習品質マネジメント検討委員会

1. はじめに

大規模言語モデルなどの基盤モデルとそれを用いた生成系 AI は昨今大きな注目と期待を集め、利用が急速に拡大している。

この「新潮流」の AI は、これまでの AI にない高度な機能を提供する一方で、従来の AI にはなかった新たな問題が生じる可能性も指摘されている。そのため、新潮流 AI を開発・利用する企業にとって、品質マネジメントの重要性がさらに高まっている。また、新潮流 AI は基盤モデルの開発者とその応用製品やサービスの開発者や提供者という分業体制に特徴がある。しかし、現状、基盤モデルを開発できるのは少数の企業や組織に限られており、応用製品やサービスの開発者や提供者にとって基盤モデルはブラックボックス化しており、品質マネジメントを行うのが難しい構成要素となっている。

機械学習品質マネジメント検討委員会は、これまでも機械学習を用いた AI の品質マネジメント手法を検討し、ガイドライン公表や国際標準化を進めてきた。さらに、この手法を新潮流 AI に適用する場合の新たな課題や、その対応を検討した。以下、それらについての現時点での見立てを述べる。新潮流の状況は変化が速いため、今後も検討を継続する必要がある。

AI や機械学習の国内関連団体からも既に新潮流を踏まえた提言やガイドラインが、2023 年春の段階で出されていたが、利用者観点での法令や倫理面の課題を考慮した内容であった。これに対し、本稿で扱うのは品質マネジメントを行うサービス開発者・提供者にとっての技術に起因する課題である。なお、サービス開発者・提供者に加えて基盤モデル開発者も含めた開発側に対する要請ないし規制は世界各地で動きがあり、後述する。

2. 概観・概要

新潮流の特徴として、1）基盤モデルに起因するサプライチェーンの複雑化と、2）生成系 AI 特有の、訓練データが原形に近い形で出力に再現される現象、の 2 つがある。以上から、AI 品質マネジメントの課題として、サプライチェーン上のステークホルダー間での責任分担と、出力におけるプライバシー・機密保持・著作物の複製の問題が生じる。

新潮流 AI について指摘されている品質問題を概観すると、応用製品やサービスの用途に

よらず常に問題となるものと、用途や利用状況によって問題になるものに大別できる。用途によらない問題については基盤モデル開発者による品質マネジメントが重要である。用途に依存する問題については、応用製品やサービスの開発者や提供者による品質マネジメントの比重が大きくなる。責任分担には特に基盤モデル開発者からの、実施した品質マネジメントに関する情報開示が必要である。また適切な責任分担のありかたについて社会合意の形成が必要である。

新潮流 AI はテキストや画像など豊かな表現力を持つ出力を生成する。出力の中に、機械学習に用いた訓練データがほぼそのまま再現される状況を完全に避けることが難しい。このため、訓練データに個人情報や営業秘密が含まれている場合、それらが出力に現れることがある。その出力の開示範囲によってはプライバシー侵害や秘密漏えいが起きる。また、著作物を訓練データとして用いることや、著作物が再現された出力の扱いについては著作権上の懸念が指摘されている。

産総研の機械学習品質マネジメント検討委員会は、これまでの取組みに加えて、上記の課題への対応にも取り組んでいく。

以降の節では、以上の内容をさらに詳しく述べる。

3. 本文書の背景と前提

3.1 本文書で扱う AI の新潮流[新想定]

ここでは、最近注目を集める AI システムとして、以下の二つの特徴を持つものを取り上げる。一つは基盤モデルに基づくこと、もう一つは生成系タスクに用いられることである。基盤モデルとは多様な用途に使うことを想定して大量のデータで訓練されたモデルで、目的とする用途向けに追加訓練を施したり、入力によって特定の用途に役立つ出力を誘導したりして用いられる。生成系タスクとは出力として文章や画像など豊かな表現力を持つデータを生成するタスクである。

基盤モデルに基づく生成系 AI の学習は従来の AI に比べて多様であり、今もなお様々な手法が考案されている。比較的以前から知られた学習の方法としては、以下の 3 つがあり、それぞれ役割の異なる入力データを用いる。最初の方法は基盤モデルの訓練である。幅広い用途をカバーする大量のデータを初期訓練に用いる。第 2 の方法は基盤モデルの調整である。目的の用途に対応させる追加訓練のことで、初期訓練と区別して調整（fine tuning）と呼ぶ。第 3 の方法はコンテクスト内学習である。運用開始後に生成系 AI に与える入力データ（プロンプト）により、タスクの具体的な詳細内容や制約条件を指示して、それに沿った出力を導く。

上記の 3 つの方法を踏まえると、以下のステークホルダーを識別できる。

基盤モデル開発者　基盤モデルの訓練を行う。

サービス開発者　訓練済み基盤モデルを調整し、また、目的のタスクを基盤モデルが実行するよう誘導するプロンプトを装備することにより、目的とする用途向けのサー

ビスを開発する。

サービス提供者　目的とする用途向けのサービスを運用する。

サービス利用者　上記サービスの直接利用者である。サービス利用者には個人と組織がある。組織の場合には組織内で使う場合と顧客向けに使う場合がある。

サービス利用者である組織の顧客　直接または間接にサービスを利用することがある（サービスを組織が個別の顧客向けに使う場合など)。

サービスを利用しない一般市民　サービス利用者やその顧客の行動によって影響を受けることがある。

行政等のガバナンス機関　サービスの社会的な影響を制御する（予防、軽減を図る、監視する、被害を補償する等)。

3.2 従来のガイドラインが扱う AI の範囲[従来想定]

これまで国内外で公表されている AI の品質やリスクのマネジメントガイドラインやフレームワークは、用途を想定して開発された機械学習モデルに基づく予測・判定系 AI を対象としている。機械学習モデルを開発する組織は、そのモデルを自ら使う場合も他者に使わせる場合も、最終用途を理解しており、その用途に合わせて品質やリスクのマネジメントを行う。また、既存のガイドラインやフレームワークは、AI システムが行うタスクとしては予測・判定・制御を想定している。これらのタスクでは出力は認識や識別結果を表すラベルや、予測値や制御の値であり、取り得る値の範囲を規定することができ、多くの場合、特定の値が持つ影響が分かりやすい。

3.3 新想定と従来想定の違い

新想定と従来想定の違いは、以下の 2 点にまとめられる。

1）基盤モデルに基づく AI：　従来想定では機械学習モデルの開発者は最終用途を理解しており、必要な品質目標を設定して品質マネジメントを行う。新想定では基盤モデル開発者は最終用途を決めることができず将来可能性のある幅広い用途向けに基盤モデルを訓練する。またサービス開発者や提供者は基盤モデル開発には直接関与せず、調整フェーズ以降でのみ品質マネジメントを行うことができる。

2）訓練データを再現しうる出力：　従来想定では出力させたい値はラベルや予測値で、あらかじめ取り得る値が分かっており、訓練データを再現する可能性はない。新想定における出力には訓練データと同等の表現力があり、訓練データを再現しうる。

4. 新想定に指摘されているリスク

新想定で問題になりうると指摘されている事項は多いが、その中には問題になるかどうかが用途に依存するものがある。それらの問題に対処する上では、問題に対処し得る立場にあるものが用途を知り得ず、用途を知るものが直接問題に対処しえない立場にある点に、より根源的な難しさが認められる。

新想定で問題になりうると指摘されている事項には以下がある。

目的とする用途に対する有用性、安全性　生成系 AI の出力が期待を満たさない場合がある。指示に従わない、指示に示した制約を満たさない、指示には沿っているが役に立たないなど。また、危険な場面での不適切な出力は安全性にも影響する。

正確性　生成系 AI は事実が求められる状況でも架空の事象をあたかも事実であるかのように出力することがある。

公平性　生成系 AI が不公平な出力を出す恐れがある。

プライバシー、秘密保持　訓練データが出力に再現される結果、訓練データに含まれる個人や組織の秘密情報が漏えいすることがある。

著作権　訓練データセットの中に著作物が入っていると著作権等の侵害が起きる恐れがある。

その他不適切な出力　用途によって不適切となる出力。攻撃的な言辞、政治的あるいは道徳的見解、犯罪幇助や扇動に関わる出力、性的あるいは暴力的な内容を含む出力などがこれにあたる。

AI セキュリティ　サービス開発者にとって基盤モデルは多くの場合外部から調達するものであり、入手以前に攻撃を受けて改ざんされている可能性がある。また生成系 AI に不適切な出力をさせる攻撃も指摘されている。

上記問題のうち、有用性、安全性、正確性、公平性、その他の不適切な出力が問題となるかどうかは、用途や利用状況に依存する。例えば上記のリスクを理解した専門家が組織内でのコントロール下で創造性を要する作業の支援に使う場合には問題にならないことがありうる。一方、プライバシーや秘密保持、著作権等の問題は、人権や法令順守に関わり、用途によらず常に問題となり得る。なお上記の問題への対策の有効性を維持するために、AI セキュリティは従来通り必要である。

4.1 著作権等の侵害問題

生成系 AI では訓練データが出力に再現されることがあるが、その際、出力には訓練データが必ずしも元の形ではなく断片的に、あるいは一部が改変された形で出てくることがある。また出典が示されないことが多い。これまでの著作権関連の法律はこのような浸み出し現象がありうることを想定し制定されてはいない。

このため、何を訓練データとして用いたかに関する情報を基盤モデル開発者が開示することは、サプライチェーンの下流で品質マネジメントを行う上で非常に重要である。

5. 世界各地の公的機関の対応状況

新想定の出現を受けて、世界各地で法的強制力を持つ規制の導入が検討されはじめている。EU では 2023 年 6 月 14 日に欧州議会が生成系 AI への言及を含む AI 法案の改定案を可決した。三者協議を経て 2024 年早期の施行を目指している。米国では連邦取引委員会(FTC)が 2023 年 7 月 13 日に Open AI に情報開示要求を送付、著作権局が 8 月 30 日に AI と著作権法に関するパブリックコメントを出す、バイデン政権が 10 月 30 日に AI の安全性

に関する新基準などを求める大統領令を公表、など、様々な動きがみられるが、次第に規制強化に傾いている。中国ではサイバースペース管理局が 2023 年 4 月 11 日に生成系 AI の管理措置に対する意見公募を開始し、8 月に制定した。

日本では 2023 年 5 月 11 日に岸田首相が G7 で国際的なルール作りを主導すると宣言し、それを受けて 10 月 30 日に G7 広島プロセスに関する首脳声明が発出され、高度な AI システムの開発を開発する組織向けの国際指針及び行動規範が公開された。

また国際標準化機関でも議論が高まっている。

6. 従来想定向け品質マネジメントの状況

これまでに公的機関が公表している AI のリスクや品質に関するマネジメントガイドラインやフレームワークとしては、産総研の「機械学習品質マネジメントガイドライン」本編の他、ドイツの Fraunhofer IAIS による AI Assessment Catalog や、米国 NIST の Artificial Intelligence Risk Management Framework がある。

これらは概ね、以下の点で一致している。

- 品質マネジメントの一次的責任は AI サービス開発者・提供者にある。
- マネジメント対象の AI システムの用途や想定利用状況などのビジネス要件を起点として技術的要件を導出し、その結果に基づいて品質目標を定め、品質評価指標や品質実現・改善手法を選んで実施する。
- 品質に対する信頼は、品質マネジメントの実施記録を残し、開示することで得られる。
- AI サービスに用いる製品・部品やソフトウェアを外部調達する場合には、AI サービス開発者・提供者は調達した製品等の品質を確保する必要がある。

これらは従来想定に基づくものである。新想定でも基本的にこの考え方は有効であり、踏襲するべきであるが、これらの項目をそのままの形で実施することは難しい。

7. 新想定に対応した品質マネジメントの見立て

基盤モデルの出現によって、基盤モデル開発者とそれに依存するサービス開発者・提供者という役割の区分が生じた。それにより、サービス開発者・提供者は機械学習要素全体の品質マネジメントを直接あるいは自らの監督の下で実施するのが難しくなった。一方で、汎用の目的に開発される基盤モデルの開発者が個別用途向けにサービス開発者・提供者と同等の品質マネジメントを実施するのも困難である。このような状況下で最終的な提供サービスの品質を管理するためには、それぞれの役割に応じた品質マネジメントの作業を基盤モデル開発者とサービス開発者・提供者の間で分担する[1]必要がある。

[1] この「作業の分担」は純粋に技術的な意味合いであり、必ずしも基盤モデル開発者が最終的なサービスの利用者に対する責任を負うことは意味しない。欧州においては、いわゆ

サービス利用者や社会全般に対しての責任は一義的には最終製品としてのサービス開発者・提供者に帰するものと考えられるが、上記のとおり技術的観点から、サービス開発者・提供者が基盤モデル開発者に依存せざるを得ない品質観点がある。例えば、学習用データの流用に伴う著作権等のリスクや、差別的な出力などのリスクを管理するためには、学習の元になったデータについての情報が欠かせない。このような観点からサービス開発者・提供者は、有償・無償を問わず可能な範囲で、自らが品質マネジメントを行うために必要十分な情報を提供している基盤モデル開発者により提供される基盤モデルを選択して用いることが望ましい。また、基盤モデル開発者が積極的にそのような情報を市場に提供することにより、社会全体での品質マネジメントの取り組みが進むことが望ましい。

8. 今後の課題

今後の課題としては、社会的には、基盤モデル開発者とサービス開発者・提供者に留まらない多数の利害関係者の間での責任やコストの分担についての合意形成がある。例えば以下の事項についての合意が必要である。

- サービス開発者・提供者が用途に応じて担うべき責任
- 基盤モデル開発者に求めるべき品質に関する情報の整理
- 基盤モデル開発者、サービス開発者・提供者それぞれについて、果たした責任の評価・明確化。具体的には実施対策に関する尺度などのコンセンサス形成
- サービス利用者やその顧客が留意すべきこと
- 監査やモニタリングの社会的制度

技術的には、必ずしも緊密な協力関係にないサプライチェーンの関係者の間で品質を確保するために必要なやり取りを効率よく効果的に行う手法の確立が挙げられる。このような技術の開発には、機械学習 AI の研究者と品質マネジメントの研究者が緊密に連携する必要がある。法令や倫理など社会的要請の専門家との連携も必要である。

産総研の機械学習品質マネジメント検討委員会は、これまでの取組みに加えて、上記の課題への対応にも取組んでいく。

る AI 法案の議論において、基盤モデル開発者に一定の責任を負わせようとするような動きが見られる一方で、同時に検討されている機械規則の議論においては、従来通りの考え方に基づき、部品としての基盤モデル提供者の責任は部品提供先との契約関係に留まり、最終製品の利用者に対する直接的な責任は最終製品提供者が負うような考え方がみられる。このような法的な責任分担の考え方は今後も各地域で議論が続くと考えられるが、技術的な依存関係と役割分担の必要性は、大きく変わらないものと考える。

9. 謝辞

本文書の作成は、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）からの受託事業の一部として、国立研究開発法人産業技術総合研究所（産総研・AIST）と大学共同利用機関法人情報・システム研究機構国立情報学研究所（NII）が、企業・大学等の有識者委員とともに構成した「機械学習品質マネジメント検討委員会」と、関心の高い有志の方々の参加を得て行った。

「新潮流への対応検討」参加メンバー（五十音順）

荒木　俊則	日本電気株式会社
磯部　祥尚	産総研
江川　尚志	産総研
越前　功	国立情報学研究所
大岩　寛	産総研
大橋　恭子	富士通株式会社
岡本　球夫	パナソニックホールディングス株式会社
小川　秀人	株式会社日立製作所
川本　裕輔	産総研
北村　弘	Community of Deep Learning Evangelists（日本電気株式会社）
金　京淑	産総研
桑島　洋	株式会社デンソー
小西　弘一	産総研
小林　健一	富士通株式会社
小宮山　英明	コニカミノルタ株式会社
妹尾　義樹	産総研
田部　尚志	日本電気株式会社
中島　震	産総研
中島　裕生	テクマトリックス株式会社
難波　孝彰	パナソニックホールディングス株式会社
浜谷　千波	アドソル日進株式会社
林谷　昌洋	日本電気株式会社
三宅　和公	住友電気工業株式会社
三宅　武司	株式会社サイバー創研
若松　直哉	日本電気株式会社
山田　敦	日本アイ・ビー・エム株式会社
Tinghui Ouyang	国立情報学研究所