2F-07

大規模言語モデルにおける説明可能なAIの法的役割

鈴木 健二
東京工業大学

概要

大規模言語モデルは，変革をもたらす高性能な技術であるが，その動作が非常に複雑である．そして，誤情報，偏見，サブリミナル操作などの潜在的な課題がある．説明可能な AI は，ブラックボックスモデルを人間が理解できるようにする技術である．2023 年 12 月，欧州 AI 法案は，基盤モデル，生成 AI，汎用目的 AI への規制を取り込んだ．大規模言語モデルについての説明可能な AI は，どのような法的役割を持つのかについて検討する．

１． はじめに

大規模言語モデルは，あらゆる分野において，広範な知識を基にした学習データを活用することにより，多大な恩恵をもたらしている．これには，人間との対話によるブレインストーミングや内容に関するアドバイス，人間が後で修正することを前提にしたドラフト作成，複雑な文書の要約や解釈，そして文章の推敲や校正などが含まれる．これらは，大規模言語モデルの登場によって可能になった効果である．

一方で，大規模言語モデルによる倫理的・法的・社会的課題も勃発した．例えば，誤情報の出力，バイアスによる差別，サブリミナル効果を利用した無意識での操作，人間が幻覚を感じるハルシネーション，ディープフェイクの拡散など，大規模言語モデルの登場によって多くの問題も生じている．

本稿は，包括的な AI 規制となる欧州 AI 法案を基に，大規模言語モデルにおける説明可能な AI の法的役割について考察する．

２． 説明可能な AI とその限界

大規模言語モデルは，複雑なアルゴリズムと広範なデータに基づいており，その出力の理由を解釈することが難しい．近年，ブラックボックスモデルと呼ばれるディープラーニングモデルの判断根拠を明示させる説明可能な AI の研究が多くなされている．説明可能な AI とは，ブラックボックスモデルを人間が理解できるようにする技術である．説明可能な AI は，人間に対して AI モデルの判断基準やプロセス透明化し，人間が AI モデルについての理解を深めるのに役立つ．ニューラルネットワークのような複雑なネットワークになるほど，その判断理由の説明が難しくなる．説明可能な AI の技術は，大まかに次の３つに分類される．(1) 線形回帰などの単純なモデルによる内在的な説明，(2) 特徴量と予測値の関係を抽出することによるポストフォック説明可能な AI，(3) 反実仮想による例示ベース手法がある．しかしながら，多様な説明手法が存在し，その判断根拠もそれぞれ異なる．また，それぞれの手法での結果を表しているが，その説明可能な AI に関する評価が難しい．なお，手法によっては再現性の問題も生じる．そして，大規模言語モデルにおける説明可能な AI に関する研究は発展途上であり，同様の問題を内在する．

３．法学における推論

Richmond らは，法学における推論について次のように整理している[1]．その分類は，(1)演繹的推論により与えられた問題に法則を適用するルールベースの法的推論，(2)論証に基づく法的推論，(3)類似の法的先行事例に基づく推論，(4)合理的な推論に基づく推論型法的推論，(5)ルールベースとケースベースの推論を組み合わせたハイブリット推論である．法学において高い説明責任，つまり透明性が要求される．彼らは，ディープラーニングモデルのような不透明な AI は，現在のレベルの説明可能な AI の技術が法領域の要件を満たさないと指摘している．

４．欧州 AI 法案

（１）大規模言語モデルへの規制

2021 年 4 月，欧州委員会は，世界初の包括的な AI 法案を発表した[2]．その AI 法案は，AI によるリスクに応じて，①原則として禁止する「許容できない AI」，②規制に基づく「ハイリ

The Legal Role of Explainable AI in Large-scale Language Models
Kenji Suzuki, Tokyo Institute of Technology

スク AI」，③透明性を求める「限定リスク AI」，④自主規制に委ねる「最小リスク AI」の 4 つに分類し，それぞれ要件が異なる．2023 年 6 月，基盤モデルや生成 AI の規制に関する条項を含んだ欧州 AI 法案を欧州議会が採択した[3]．2023 年 12 月，欧州議会，欧州委員会，EU 理事会の三者は，基盤モデル、生成 AI に加えて汎用目的 AI も含んだ欧州 AI 法案について政治的合意に至った．

（2） 欧州 AI 法案での説明可能な AI の役割

大規模言語モデルが欧州 AI 法案に含まれる前の先行研究であるが，Panigutti らは，欧州 AI 法案が透明性に関して説明可能な AI を利用することを義務としているものでもなく，ブラックボックスモデルの利用を禁止しているものでもないと解釈している[4]．欧州 AI 法案は，文書化を含む透明性と人間の監視に焦点を当てることで目標を達成することを目的としている．

大規模言語モデルについて欧州 AI 法案は，「独立した専門家の関与によるモデル評価，文書化された分析，概念化，設計，開発中の広範なテストなどの適切な方法を通じて評価された性能，予測可能性，解釈可能性，適格性，安全性，サイバーセキュリティの適切なレベルを求めている」（欧州 AI 法案 28 条 2 項(a)）．ここで述べられている「解釈可能性」とは何を示すのだろうか．

５．法的役割の検討

このような状況を鑑み，大規模言語モデルにおける説明可能な AI の法的役割について検討する．欧州 AI 法案は，説明可能な AI を必ずしも要求していない[4]．2023 年 12 月に欧州 AI に取り込まれた大規模言語モデルへ規定された「解釈可能性」（欧州 AI 法案 28 条 2 項(a)）ついても同様に適用できると考える．また，説明可能な AI は，法的文脈では法領域における要件を満たさない[1]．法は，推論の過程において，どのようにして意思決定に至ったのかを明確に求めており，現在の説明可能な AI は十分な技術ではない．AI が法的決定やプロセスにどのように組み込まれているかにもよるが，大規模言語モデルでの説明可能な AI は，法的な役割としては透明性を高めるための補助的なものに過ぎず，法的な要求や要件までも満たすものではない．

今後，技術が進歩して説明可能な AI が法的な役割を果たすためにはどのようになればよいのだろうか．ルールベースの推論システムのように，人間が判断根拠を理解しやすいものは受け入れられるだろうが，高度な大規模言語モデルのようなものはその複雑な動作のため，法的に意味のあるレベルの説明手法が確立されるのはかなりの困難を伴う．また，AI の予測や説明可能な AI 自体の信頼性の問題もあり，そのような誤差や誤りを明確にすることも求められる．そして，説明可能な AI は，そのモデル自体についての判断根拠を提示するが，これは説明責任とは異なることに着目しなければならない．AI の開発において様々なレイヤーでのバリューチェーンが存在する．データの収集，データの前処理，ディープラーニングモデルの学習，アプリケーションへの実装，製品・サービスの運用などは，1 社で必ずしも完結できるものではない．ディープラーニングモデルでの問題が起きた場合，責任の所在がどこにあるのかを明確にしなければならない．これは，説明責任であり，説明可能な AI の技術とは異なるものである．

６．まとめ

欧州 AI 法案での大規模言語モデルへの規制動向に基づき，大規模言語モデルにおける説明可能な AI の法的役割について検討した．その結果，大規模言語モデルの説明可能な AI は，法的な役割として必ずしも要求されているものではない．また，欧州 AI 法案は，ブラックボックスモデルである大規模言語モデルを禁止するものでもない．

参考文献

[1] Richmond,K.M., Muddamsetty,S.M., Gammeltoft-Hansen,T,Olsen,H.P and Moeslund,T.B.: Explainable AI and Law: An Evidential Survey, Digital Society, Volume 3, Article Number 1, (2024).

[2] European Commission: Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, (2021 年 4 月 21 日).

[3] European Parliament: DRAFT Compromise Amendments - European Parliament on the Draft Report - Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206-C9 0146/2021-2021/0106(COD)) (2023 年 5 月 16 日).

[4]Panigutti,C.,Hamon,R.,Hupont,I.,Llorca,D.F.,Yela,D.F.,Junklewitz,H.,Scalzo S.,Mazzini,G., Sanchez,I.,Garrido and J.S.,Gomez,E.: The role of explainable AI in the context of the AI Act, In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, June 12–15, (2023).