日刊工業新聞
年月日 19 07 05 ページ 13 NO.

ICT 世界の潮流 PART7 4

AIの倫理ガイドラインと透明性の原則㊦

国際社会経済研究所（NECグループ）主幹研究員
小泉 雄介

2通りの透明性確保

前回紹介した欧州委員会の人工知能（AI）倫理ガイドラインでは、7要件の一つとして「透明性」が挙げられている。透明性は、AIシステムが利用者から信頼を獲得し、我々の社会に不可欠な存在として受容されていく上で、非常に重要な要件である。

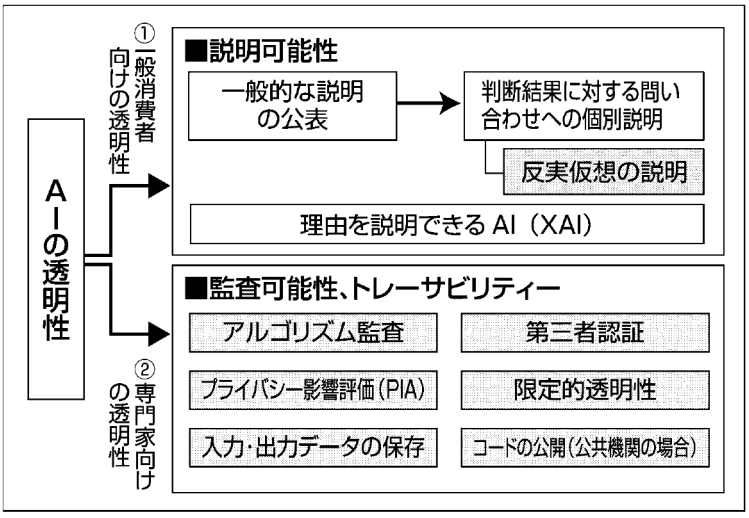
AIの問題点

機械学習に基づくAIの問題点として指摘されるものの一つに、AIのブラックボックス化がある。例えば、自分の医療データを機械学習型のAIで分析してもらい、心臓病にかかるリスクが高いという結果を受け取った場合、なぜそのような判断結果になったかの根拠を知りたいだろう。しかし、機械学習型AIはあくまで統計的パターンに従って結果を返してくるため、その判断に至った理由や根拠を人間の言葉で表現することは難しいとされる。

このようにAIによるデータ分析の透明性を確保することは大きな課題となっているが、AIを開発・利用する企業の取り組みとしては、二つの方向性がある。①一般消費者向けの透明性と、②専門家向けの透明性だ。①は説明可能性を追求するものであり、消費者向けに分かりやすい説明を工夫し、なるべくブラックボックスを緩和しようとする。②はアルゴリズム監査、認証など、第三者の専門家にAIの正当性を保証してもらう方向性である。

合理的な説明

①では、必ずしもAIのコードや技術的詳細（全てのパラメーター、その重み付けなど）を開示することは重視されず、むしろ消費者の「なぜそのような判断結果になったのか」という問いに対して、本人が理解できる形式で合理的な回答を行う「説明可能性」が重要となる。この説明可能性は、AIの判断プロセスに関する一般的な説明（判断に影響を与える主なデータ、それらの情報源など）の公表と、ある判断結果（特に不利な結果）を受けた消費者に対する個別説明という2段階から成る。



「反実仮想」

こうした判断結果の根拠や理由を説明する手法の一つに、「反実仮想」の説明がある。これは「あなたのデータがどのような条件を満たしていれば、良い判断結果を得ることができたのか」を説明するものだ。例えば消費者のローン申請が拒否された場合、年収がいくらならば申請が通ったのかを教える手法ということになる。理由を説明できるAI（XAI）は、このような説明可能性を底上げする取り組みと言える。

②では、消費者の代わりにAIの専門家が透明性を保証する。米国では既にAIの透明性・公平性・正確性などについて監査を行い、ボランタリーな認証を行う企業が存在する。限定的透明性という考え方もあり、これは消費者が監査報告書などの説明を理解できなかったとしても、データ保護監督機関が理解できればよいというものだ。

プライバシー影響評価（PIA）はシステムの設計段階でネガティブな影響を特定し低減するための手法で、特にバイアスの低減に役立つだろう。またAIの入出力データの保存は、何らかのエラーが発生した際に検証でき、将来的なエラー防止にも寄与するだろう。（金曜日に掲載）

無断転載・複写禁止 ㈱日刊工業新聞社