日刊工業新聞
年月日 19 06 21 ページ 15 NO.

ICT 世界の潮流 PART7 3

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

害悪回避へ倫理原則

人工知能（AI）技術は、個人にも企業や社会にとっても多大なベネフィット（便益）をもたらすことが期待されているが、同時に個人や社会に対して新たな害悪を及ぼすリスクについても懸念されている。

人間中心のAI

AI開発や利用の諸原則については、既に国内外の政府機関などがガイドライン類を発行している。わが国では総務省のAIネットワーク社会推進会議が6月にAI利活用ガイドライン案を、内閣官房の「人間中心のAI社会原則会議」が3月にAI社会原則を公表している。海外では、欧州委員会のAIハイレベル専門家グループが4月に「信頼できるAIのための倫理ガイドライン」を、英国でも2018年3月に議会の特別委員会がAIの包括的原則を公表しており、代表的な取り組みとして17年のアシロマAI原則がある。また、経済協力開発機構（OECD）は19年5月にAIに関するOECD原則を採択した。なお、つくば市で6月上旬に開催された主要20カ国・地域（G20）貿易・デジタル経済相会合では、このOECD原則と同じ内容のAI原則を含む閣僚声明が採択された。

7つの要件

欧州委員会のAI倫理ガイドラインでは、①「人間の自律の尊重の原則」「害悪防止の原則」「公平性の原則」「理解可能性の原則」の四つから成る。これらの倫理原則が挙げられている。さらに、これら7要件をAIシステム向けの具体的な要件にマッピングするために、③と④は「人間の自律の尊重の原則」「害悪防止の原則」「公平性の原則」「理解可能性の原則」の四つから成る。これらの倫理原則が挙げられている。さらに、これら7要件をAIシステム向けの具体的な要件にマッピングするために、③と④の技術的手段と非技術的手段が挙げられ、技術的手段には設計段階からの倫理、説明できるAI（XAI）など、非技術的手段には法規制、行動規範、標準化、認証などが含まれている。また⑤は、上記7つの要件への適合を評価するためのパイロット版のリストである。

欧州委員会AI倫理ガイドラインの7要件

要　件	サブカテゴリー
1.人間の主体性と監督	・基本的人権 ・人間の主体性 ・人間による監督
2.技術的な堅牢性と安全性	・攻撃からの回復力とセキュリティー ・予備プランと全般的安全性 ・正確性 ・確実性と再現性
3.プライバシーとデータガバナンス	・プライバシーとデータ保護 ・データの品質と完全性 ・データへのアクセス
4.透明性	・トレーサビリティー ・説明可能性 ・コミュニケーション
5.多様性、非差別、公平性	・不公平なバイアスの回避 ・アクセシビリティーとユニバーサルデザイン ・ステークホルダーの参加
6.社会および環境の福祉	・持続可能で環境にやさしいAI ・社会的影響 ・社会と民主主義
7.説明責任（アカウンタビリティー）	・監査可能性 ・ネガティブな影響の最小化と報告 ・トレードオフ ・救済手段

AIの透明性

本稿のテーマであるAIの透明性に関わるものとして、「理解可能性の原則」では、「なぜシステムが特定のアウトプットや判断を行ったのか、またどのインプットの組み合わせがアウトプットに寄与したかについての説明は、常に可能なわけではない（いわゆるブラックボックス問題）。それらのケースでは、他の理解可能性の手法、すなわちトレーサビリティー、監査可能性、システム能力に関する透明性などが必要とされるかもしれない」と言われている。AIの透明性については次回取り上げたい。（金曜日に掲載）