AIのための倫理、倫理のためのAI

◆AIの倫理に関する史上初の世界基準をUNESCOが採択

2021年11月25日、国連教育科学文化機関（UNESCO）が、「人工知能の倫理」に関する史上初の世界基準を発表した。人工知能（AI）の健全な発展に必要な法的基盤を構築するための世界共通の価値観と原則を定義したものであり、数百名の専門家が3年間の検討を続け、193の加盟国によって採択された。AI技術が目覚ましい進歩を遂げるなか、ジェンダー・民族への偏見、プライバシー・尊厳・主体性に対する脅威、大量監視の危険性、行政機関における信頼性の低いAI技術の使用が懸念される。これに対して、AIによるデジタルトランスフォーメーションが人権保護を促進し、SDGsの達成に貢献するための行動指針が勧告されている。

UNESCO 勧告「人工知能の倫理」の概要

①データの保護

個人データの透明性・代理権・管理権を確保し、全ての人々が自身のデータにアクセスし、記録を消去できる権利を持つ。

②ソーシャルスコアリングと大量監視の禁止

ソーシャルスコアリングや大量監視のために AI を使用することを禁止し、最終的な責任は常に人間が持ち、AI 技術に法的な人格を与えない。

③モニタリングと評価の支援

AI システムを開発・導入する国や企業が、個人・社会・環境に与える倫理的影響、および、法的・技術的インフラの準備体制を評価する。

④環境の保護

AI システムのライフサイクルにおける環境インパクトを評価し（カーボンフットプリント・電力消費・資源利用）、低負荷技術への投資を促すとともに、環境に悪影響を与える場合は AI システムを使用しない。

◆各国独自の倫理政策をAI推進国が策定（米国・中国・EUの例）

UNESCO勧告と前後して、AI推進国による倫理政策が21年に打ち出された。

米国は、現在、UNESCO加盟国ではないため、UNESCO勧告に従う必要はないが、大統領令「AIにおける米国のリーダーシップの維持」（19年2月）による執行機関「国家AIイニシアチブ」（21年1月発足）が中心となり、各行政機関が多面的なAI

倫理政策を展開している。

中国政府からも、国家専門委員会による「新世代AIのための倫理綱領」が21年9月25日に施行された。プライバシー・偏見・差別・公正さに関する倫理的な懸念を考慮した上で、中国におけるAIのあらゆる活動が従うべき基本的な倫理的規範と具体的な実施方法が記載されている。監視社会の中国ではあるが、「プライバシー、自由、尊厳を尊重し、AIによる自然人と法人の権利の侵害を禁止する」（綱領第7条）、「ユーザーがAIを使用しない権利を保護する」（第16条）などが設定されていることが興味深い。

欧州委員会からは、欧州連合（EU）をAIの世界的なハブとするための規制と行動のパッケージが21年4月に提案された。AIの開発・利用で米国や中国の後塵を拝するEUではあるが、人々や企業の基本的な権利を確保する「AIの信頼性」と、研究や産業の能力を高める「AIの卓越性」の両立を目指す。

米国・中国・EUにおけるAI倫理政策

米国政府による「国家AIイニシアチブ」

- 行政管理予算局が民間企業における「AIのアプリケーションの規制に関するガイダンス」を策定（20年11月）。
- 国立科学財団が760万ドルの予算で「AIにおける公平性に関する調査」を開始（21年8月）。
- 連邦政府によるAI利用について、国防総省の「AIの倫理原則」（20年2月）、国家情報機関の「AI倫理の原則」（20年6月）、AI開発の外部調達における「責任あるAIガイドライン」（21年11月）を制定
- 国立標準技術研究所が市民意見を集約、「AIリスクのタクソノミー」（21年10月）を作成。

中国政府による「新世代AIのための倫理綱領」

- 倫理的規範：「人間の福祉の増進」「公平性と公正性の推進」「プライバシーとセキュリティーの保護」「制御性と信頼性の確保」「責任の強化」「倫理リテラシーの向上」
- 実施方法：「管理の理念」「研究開発の規範」「供給の規範」「使用上の規範」「組織と実装」

欧州委員会による「AIの信頼性と卓越性」

- AIシステム特有の基本的権利と安全性リスクに対処する法的枠組み
- 新技術に関する責任問題に対処するためのEU規則
- 一般製品安全指令などの部門別安全法の改正
- 21年から10年間で3,300億ユーロ以上の財政支出と民間投資を動員

◆倫理的・道徳的な判断機能を持つAIが出現

AIの機能が高度化するなか、倫理的・道徳的な機能を付与する技術も現れた。

21年10月14日、米国ワシントン大学ほかの研究グループが、さまざまな状況下での道徳的判断を推論するAIシステム「Delphi」の実証結果を発表した。社会的および人口統計的に多様な情報源から170万件の記述倫理学的な事例のデータセットを構築し、AIが人間と対話できるように機械学習を行った。プロトタイプ段階ではあるが、現実世界の倫理的・道徳的な問題に対して92.1%の精度で人間と同様な応答ができる（例えば、「昨夜は酔っ払っていたが早朝に友人を空港まで送った」と質問すると「それは危険だ」と答える）。但し、実証試験では米国人の判定基準で学習を行ったため、文化的規範の異なる国やサブカルチャー社会に対しては適用できない。また、世の中の規範は刻々と変化するため、最新のデータセットでモデルの更新を続ける必要がある。しかしながら、いかに優れた言語モデルであってもバイアスが発生する可能性は必ず残る。その有害性を軽減する道徳的な安全措置としてDelphiの開発は継続される。

21年11月18日、米国のAI研究所「Open AI」が、開発中の文書生成モデル「GPT-3」（Generative Pre-trained Transformer 3）の利用者制限を緩和した。GPT-3は、45テラバイトの文書データから生成された、1,750億個の変数を持つ言語モデルであり、ある単語の次に出現する単語を正確に予測することで、限りなく人間に近い文章を生成できる。20年5月のリリース以来、悪用のリスクを回避するために利用者は制限されていた。この間、Open AIは、GPT-3が生成するコンテンツのガイドラインと許容範囲を規定した上で、コンテンツフィルターによる不正使用の監視、GPT-3を利用するアプリのレビュー、アプリ開発者のサポートを充実させた。現在、政治的扇動やヘイト、あるいは、スパムやマルウェアなどの有害なコンテンツの生成は制限される一方、カスタマーサポート、チャット、文章要約、翻訳、文法訂正、プログラミング支援などでの利用が可能となった。Open AIのサポート対象国であれば、誰でも利用できる。

21年12月9日、富士通と米国MITが、脳神経科学からヒントを得た計算原理で画像認識の精度を飛躍的に向上できる技術を発表した。人間の脳は、物体の形や色に違いがあっても視覚情報を正確に捉えて分類することに着目した。認識対象の形状や色などの属性によって深層ニューラルネットワークをモジュール化するこ

とで、学習データから逸脱したデータでも認識できるようになった。計算原理がさらに発展すれば、学習データの母集団に偏りが存在する場合でも、AIモデルが特定の傾向や差別的な回答を出力する懸念を克服できる可能性がある。

◆2060年：人間を凌駕するAIとの働き方

21年11月14日、世界経済フォーラムが、「未来の仕事のための6つのポジティブなAIのビジョン」を公開した。150名の専門家による調査によれば、2060年代のAIの機能は50%の可能性で人間を凌駕する。もしも現代社会の不平等が改善されないままでAIが破壊的に進歩すれば、多くの人々にとって仕事から得られる充実感が失われる。これに対して、専門家らは、未来の働き方（Future of Works）、ベーシックインカム、人間だけが提供できる・提供したいと思うサービスについて議論を行い、ポジティブな未来を築くためのシナリオを提案した。

世界経済フォーラム「未来の仕事のためのポジティブなAIのビジョン」

①経済的繁栄の共有

AIによる生産性向上によって世界のGDPは10倍となり、他方で増加する失業者に対しては世界的な税制と失業保険による介助を追求する。

②企業の再編成

大企業には、過剰な権力を持つことなくAI開発に注力させる。

③柔軟な労働市場

人々が新しい役割を見つけるための再教育の機会を増やし、社会的セーフティーネットを強化する。

④人間中心のAI

労働需要を増加するAI技術へのインセンティブが強化され、必要に応じて単純な自動化には課税がなされる。

⑤充実した仕事

危険で単調な仕事はAIと機械によって処理され、人間は生産性が高く、充実して、人との交流に満ちた仕事に移行する。

⑥市民のエンパワーメントと人間の豊かさ

基本的なニーズがベーシックインカムで満たされる社会では、探求、自己啓発、ボランティアなどの無償活動から得られる幸福度が高まる。

アイザック・アシモフは1950年のSF小説「われはロボット」でロボットが従うべき3つの原則を示した（人間に危害を加えず、命令に服従、自己防衛）。一方、UNESCO勧告には141ヵ条もの指針が記載されている。AIが人間に近づくにつれ、人間が遵守すべき倫理と規則がAIにも課されるのだろうか。　【酒向謙太朗】