世界最先端のAIシンポジウム AIに求められる説明責任と倫理、AI研究者がシリコンバレーで熱い議論

「Make AI Trustworthy! Explainable and Ethical AI for Everyone」（説明可能で倫理的な信頼に足るAIを求めて）—Fujitsu Laboratories Advanced Technology Symposium（FLATS）2018現地報告

FUJITSU JOURNAL / 2018年11月16日

現在、最も注目されているテクノロジー分野はAI（人工知能）と言えるでしょう。AIは1960年代から研究されてきた分野ですが、ここ数年間の進化とその適用には目覚ましいものがあり、今、二つの先端分野で多くのAI研究者が研究活動に取り組んでいます。

AIの実用化と広がりが「説明責任」と「倫理」をAIに求める

一つは技術の先端です。具体的には、人間の神経構造を模倣したニューラルネットワークで学習を行ってアルゴリズムを生み出すことで、プログラム不要で回答を出すディープラーニングです。ディープラーニングは、現時点においては、特定の分野で利用される「狭いAI」ですが、技術的進化を重ねていけば、近未来において「汎用AI」も実現されるのではと期待されています。

もう一つは、倫理的側面からAIを捉えることです。AIはいかにあるべきか、人間との関係をどう築くべきかを考察する取り組みで、オックスフォード大学などの世界の先端的な研究所で、幅広い分野からのアプローチに基づいて進められています。この取り組みは、汎用AIの出現に備えた動きであると同時に、実用段階に入ってビジネス導入が進んでいる狭いAIにおいても重要な課題として捉えられています。

こうしたAIの最先端の動きをカバーするシンポジウムが、2018年10月9日、富士通研究所が主催してシリコンバレーのサンタクララで開催されました。シンポジウムの名称は『Fujitsu Laboratories Advanced Technology Symposium（FLATS）2018』。富士通研究所が毎年テーマを決めて開催しているもので、今年のテーマとして掲げられたのが「Make AI Trustworthy! Explainable and Ethical AI for Everyone」（説明可能で倫理的な信頼に足るAIを求めて）でした。

「Fujitsu Laboratories Advanced Technology Symposium 2018」

富士通のAI技術をシリコンバレーで世界のAI研究者に問う

AIの世界で倫理が注目され始めた背景には、利点と共にAIが悪用されるケースが増えてきたことがあります。ディープラーニングによる映像加工により、実はオバマ大統領自身が話したこともない内容を口にしている演説ビデオが偽造されたケースもあります。また、米国で開発された再犯率を推定するAIにおいて、白人よりアフリカ系アメリカ人の方が可能性を高く見積もられる傾向があったことなども知られています。前者は意図的な悪用ですが、後者は十分な評価なしに既存のデータを学習させたためにAIが不当なバイアスを持ってしまったことによるものです。

後者のような"偏ったバイアスを抱えるAI"は、ビジネスでのAI利用においても問題になりつつあります。また、それと同時にAIがどのように回答を導き出すのかがブラックボックス化して人間には理解不能になっていることも、ビジネスや医療の分野などで問題視され、AI導入に対する抵抗感の一つとなってきたことも事実です。

富士通研究所
代表取締役社長
佐々木 繁

富士通研究所の佐々木繁代表取締役社長は、今回のシンポジウムのテーマについて、「AI への新たなアプローチである富士通の『説明可能なAI』への評価を、ICT研究開発の中心地であるシリコンバレーで問いたかった」と述べています。

説明可能なAI とは、AIが下した判断を"見える化"することと言えます。つまり、「AIがなぜその推定にいたったのか？」推定結果の理由や根拠を人が説明できるようになるのです。今回のシンポジウムでは、富士通研究所が開発した説明可能なAIを実現する技術として、人とモノの関係性を表現するグラフ構造のデータを解析する独自の機械学習手法「Deep Tensor®（ディープテンソル）」と、学術論文など専門的な知識を蓄積したナレッジグラフを融合したシステムが、シリコンバレーのAI研究者に紹介されました。Deep Tensor®とナレッジグラフを結びつけることで、AIの回答の背後にある推定理由や学術的根拠が提示されます。この手法は医療や金融などのミッションクリティカルな分野に有用で、佐々木社長は「説明性という企業の社会的責任を果たせる技術」と位置づけました。

富士通研究所
人工知能研究所
所長
岡本 青史

興味深いのは、富士通研究所の人工知能研究所の岡本青史所長がDeep Tensor®とナレッジグラフを結びつける技術について「結果の出し方について説明できるということは、そのプロセスの中にAIの倫理が含まれている」と指摘していることです。これは、技術と倫理という、AIの二つの先端分野に重なる世界でも数少ないアプローチと見ることができます。

世界中で本格化するAIの倫理の研究

AIの倫理については、オックスフォード大学、ケンブリッジ大学、スタンフォード大学、カリフォルニア大学バークレー校など世界のアカデミア機関で研究を進める組織が設けられているほか、私設の研究所やAI研究者らが独自に作ったFuture of Life Institute（FLI）などのグループも存在しています。いずれもAIの長期的な影響を研究したり、人に友好的なAIの開発をどう進めるべきかを議論したり、様々なバックグラウンドを持つ研究者らが関わっています。

オックスフォード大学
インターネット研究所 デジタル倫理ラボ
副所長
マリアロザリア・タデオ博士

例えばオックスフォード大学インターネット研究所内にあるデジタル倫理ラボでは、AIやデジタル革新によって引き起こされる倫理的な課題について研究しています。
この取り組みは、新しいテクノロジー利用におけるガイドラインを策定することを目的としています。同ラボの副所長であるマリアロザリア・タデオ博士は、AIの倫理はすでに15年前から課題になっているといいます。最近は、コンピュータ処理能力とデータが莫大に増えたためAIがあらゆる分野に浸透し、企業をはじめ、医師や法律家、政策決定者らがAIの倫理について大きな関心を持つようになりました。

AIに課せられる倫理の特徴とは

AIをどう統制すべきかの倫理分析は、世界中で進められています。研究を進める中で、倫理分析は道徳や文化的価値に依るところが大きく、コンテキスト（文脈）によって変化するため、どう

すれば世界中で通用するAIの倫理的原則を定義することができるのかについて議論されるようになりました。タデオ博士はデジタル倫理ラボで行われたジョディ・コウルズ、およびルチアーノ・フロリディ両氏による研究成果から、恩恵、非有害性、自律性、正義という生命倫理学における4つの原則がAIの倫理分析の基礎となると考えています。

タデオ博士は、「AIの倫理を考える際には、常に様々な要素のバランスをとることが必要」といいます。例えば、個人データへのアクセスによってアルツハイマー研究をさらに進展させることが可能になるかもしれません。しかしこの場合、研究には個人データを含む医療関連の膨大なデータへのアクセスを必要とするため、プライバシー保護と医療の進化とのトレードオフが求められます。同様に、プライバシーとセキュリティーのバランスを考慮しなければならないようなケースもあるでしょう。

AIによる恩恵を知ることで、議論の本質を捉える

AIは、革新的な技術です。電力や機械工学など他の革新的技術と同様、私たちの社会の構造に組み込まれ、社会のダイナミクスを再構築し、従来の慣習を破壊し、それによって大きな転換を起こすでしょう。だからこそAIには、新たな基礎技術としての独自の倫理的フレームワークが必要となります。

AIによる転換は、私たちのまわりの環境や社会、人類の発展についても倫理的問題を提起します。工場から道路、スマートシティーに至るまで、AIはAIがルーティーンとして機能していくための環境の再設計を促します。その設計の際に、基本となる価値は何か、それによって生み出される恩恵はどんなものか、AIフレンドリーな世界に変換していくことで生じる見えないリスクは何か、これらを理解することは必要不可欠と言えるでしょう。

同博士は「AIの倫理の研究には、ゴールはない」と語ります。これは結局、人間理解を深化させていく永遠のプロセスであり、技術の発展によって何度も改訂を迫られるものだからです。

人間と共存するAIに求められる3つの原則

カリフォルニア大学バークレー校のヒューマン・コンパティブルAIセンター（CHAI）も、人間と共存するAI研究を行なっている組織です。センターでは、コンピュータ科学、AI、ロボティクス、政治学、ビジネス、哲学、社内学など様々なバックグラウンドを持つ同学内外の教授陣が数10人集まり、AIを人間にとって恩恵のあるシステムとして開発する活動を推進しようとしています。

CHAIのエグゼクティブ・ディレクターのマーク・ニッツバーグ氏によると、同センターでは人間と共存するAIの3つの原則を次のように定義しています。

1. AIは人間の嗜好や人間にとっての恩恵を最大限実現することだけが目的である。
2. AIは当初はその恩恵が何か知らないが、それを再考し続ける。
3. その恩恵に関する情報は、人間の振る舞いの中にある。

カリフォルニア大学バークレー校
ヒューマン・コンパティブルAIセンター
エグゼクティブ・ディレクター
マーク・ニッツバーグ氏

現在のAI開発では、作業を効率化してスピードを出すことや利益を最大化することが目的とされて、必ずしも人間にとっての恩恵を優先させてはいません。その結果、当初はいいものに見えても、時間が経つにつれAIによる弊害が出てくることもあります。肝心の人間が疲弊したり、人間がAIに従属したりするような事態も起こるかもしれません。

CHAIでは、人間にとっての恩恵をAIのプログラムの中に組み込むべきであるという考えを持っています。例えば、ロボットを利用した研究では、A点からB点へカップを移動させるロボットアームが、カップを落とす危険性を避けようとする人間の教示によって、少し遠いルートでも低い位置で移動させるように学習した例があります。通常ならば最短距離を選ぶところを、人間の振る舞いから人間側の恩恵を組み込んだと言えます。

人間との共存に欠かせない価値観の一致と学習の継続

CHAIでのアプローチの特徴には、興味深い点が二つあります。一つは、CHAIでは人間と共存するAIに求める倫理として、「価値観を合わせること」を重視していることです。正しいとか正しくないというよりも、もっと使い手の人間が持つ嗜好に直接的に合致させることによって共存が実現されるという考え方です。

もう一つは、AIには固定化した目的をもたせず、人間側の嗜好や恩恵が証明可能なものとなれば、それらを学習して統合していくようにするということです。AIは、それ自体の目的を持た

ず、常に目的については不確実な状態であり、また人間の振る舞いを学習することによってその目的が変化していくのです。

ニッツバーグ氏は、汎用AI時代には安全性を考える必要が出てくると語ります。AIがAIを作るようなことも可能になり、人間がペット化されてしまうリスクも出てくるかもしれません。それに備えて、AIが人間と共存することを目的とし、人間を超えることがないよう、人間側の嗜好や恩恵の最大化を目的として設定するというわけです。現在のAI開発を根本から捉え直す重要なアプローチではないでしょうか。

AIを人間にとって意味あるものにするには、技術、哲学、社会など多様な領域からの考察が求められます。AIの発展と同時に、この未踏の分野でも様々な議論と実験が繰り広げられているのです。

著者情報

林哲史

日経BP総研　クリーンテック ラボ　主席研究員

1985年東北大学工学部卒業、同年日経BPに入社。「日経データプロ」「日経コミュニケーション」「日経NETWORK」の記者・副編集長として、通信/情報処理関連の先端技術、標準化/製品化動向を取材・執筆。2002年「日経バイト」編集長、2005年「日経NETWORK」編集長、2007年「日経コミュニケーション」編集長を歴任。「ITpro」、「日経SYSTEMS」、「ITpro」、「Tech-On!」、「日経エレクトロニクス」、「日経ものづくり」、「日経Automotive」等の発行人を経て、2014年1月に海外事業本部長。2015年9月より現職。2016年8月より日本経済新聞電子版にて連載コラム「自動運転が作る未来」を執筆。2016年12月に「世界自動運転開発プロジェクト総覧」、2017年12月に「世界自動運転/コネクテッドカー開発総覧」を発行。2011年よりCEATECアワード審査委員。

（編集協力：　瀧口範子）

FUJITSU JOURNAL　/　2018年11月16日