サイバーセキュリティにおけるAIの倫理課題

FUJITSU JOURNAL / 2020年3月2日

人工知能（AI）を取り巻く倫理は、富士通にとって極めて重要な課題です。そこで、人間中心のイノベーションを推し進めるために、その可能性を制約せず、且つAIを適切に応用できる方法を見出すために、欧州富士通研究所（Fujitsu Laboratories of Europe）が主要な役割を果たしています。

富士通は、この課題をさらに検討していくために、オックスフォード大学インターネット研究所のデジタル倫理研究所（Digital Ethics Lab（DELab））と協力し、2019年11月8日、「信頼されるAI（trustworthy AI）」に関する国際ワークショップを開催しました。

サイバーセキュリティの確保に向けたAIの適用は倫理遵守が前提

オックスフォード大学セントクロスカレッジで開催されたワークショップには、世界をリードする学界、政策立案当局、民間企業の専門家が一堂に会し、「サイバーセキュリティにおいて、私たちは信頼されるAIを開発できるだろうか」という極めて重要な問題について討論しました。

左からDr. Taddeo氏、Dr.中田氏、稲越氏、Dr. Naseer氏、Dr. Agrafiotis氏

＜ワークショップ参加者＞

- Ioannis Agrafiotis（欧州ネットワーク情報セキュリティ機関（ENISA）ネットワーク情報セキュリティ担当官 博士）
- Anna Jobin（チューリッヒ工科大学健康倫理・政策研究所 博士）
- Nathalie Smuha（ルーヴァン・カトリック大学 国際・欧州法学科）
- Mariarosaria Taddeo（オックスフォード大学インターネット研究所 博士）
- Paul Timmers（非営利シンクタンク 欧州政策センター）
- 中田 恒夫（富士通研究所 博士）
- その他、英国国立サイバーセキュリティセンターの代表者の方々

ワークショップでは、サイバーセキュリティ面について徹底的に考察するために、サイバー空間でAIが誤用、乱用された実例に焦点を当てました。さらに、データと機器の間の信頼性やモノと

モノをつなぐインターネット（IoT）の先端で発生する運用リスク、また、機器同士が自律的に通信を行うM2Mで発生する制御リスクについて討論しました。

特に、サイバーセキュリティにおけるAIの標準化及び認証手続に体現されるべき主要な倫理原則については、明確に定義された所見がいくつか提起されました。

また、専門家会合はAIの活用や誤用に関する説明責任にならんで、AIの透明性も基本原則とすること。さらに、情報システム及び情報通信ネットワークの安全性及び信頼性を確保するためにサイバーセキュリティにAIを用いる場合には、倫理原則が守られるように基準設定を行うべきであることに合意しました。

また、「信頼できるAIに関するEU倫理ガイドライン」の有効性、及びそれを裏打ちするCowls & Floridiによって定義された善行（Beneficence）、非悪意（Non-maleficence）、自律性（Autonomy）、正義（Justice）、説明可能性（Explicability）という基本原則についてもおおむね合意が得られました。

評価が分かれた「信頼されるAI」の妥当性

ワークショップでは、「信頼度」についても活発に討議されました。パネリストはAIに対する信頼度がサイバーセキュリティへのAI導入支援において大きな要因となるという点で同意しました。しかし「サイバーセキュリティにおいて信頼できるAI」の評価の妥当性については合意が得られませんでした。

特に一部のパネリストから、AIアプリケーションはサイバーセキュリティの面で必ずしも信頼性が保証されていない（希少事象や意図的な攻撃に対してAIが常に適切に対応することは不可能である）ため、コンピュータやネットワークのセキュリティ確保のために必要なオペレーションをAIに任せるには、現時点では何らかの制御構造が必要であるという意見がありました。つまり、AIの標準化及び認証手続には、これらの問題に配慮する必要があるということです。

考慮すべき3つの問題とは

ワークショップの討論では、主に次の3つの問題が浮き彫りになりました。

- AIシステムの堅牢性（robustness）は、システム設計とトレーニングとの関係性とならんで、展開後AIに与えられるデータや他のエージェントとの間に起きる相互作用に強く影響される。そのため、サイバーセキュリティにおけるAIシステムの堅牢性予測には問題が残る。
- 「信頼できるAIに関するEU倫理ガイドライン（EU Guidelines on Trustworthy AI）」は、AIの設計と利用に関して、価値のある高い基準原則を提示している。しかし、サイバーセキュリテ

ィ利用でみられる個々のAI事例に適用するには、更に細かく考察する必要がある。

- AIシステムがサイバーセキュリティのタスクを実行する場合、その堅牢性に焦点を置いた標準化及び認証手続は、自ら学習したり、動的に変化したりするAIシステムの特性が考慮されている範囲のみで有効といえる。つまり、設計から開発までの全段階を通した監視制御の形態を構想することが必要となる。

富士通はEC AIハイレベル専門家会合による「信頼できるAIに関する倫理ガイドライン 'Guidelines for Trustworthy AI'」（2019年4月発行）の策定に貢献してきました。これはEU全体のAI戦略に大きな影響を与えるもので、今回のワークショップはこの貢献の上に築かれています。

サイバーセキュリティを確保するにあたり、AI導入にメリットがあることは間違いありません。しかし、今回のワークショップではAIを適切且つ倫理的に利用するためには、さらに多くの作業が必要であることが明らかになりました。

※この記事はFujitsu Blogに掲載された「Ethical Implications of Artificial Intelligence in Cybersecurity」の抄訳です。
※本記事の文中のリンクは英語ページに推移します

FUJITSU JOURNAL / 2020年3月2日