2022年12月26日

AIリスク・ガバナンスの世界的合意形成にむけて。富士通の取り組み

AI活用に伴う課題が2010年後半から現実的な社会問題として認識されるようになっています。

GPAIと呼ばれる組織はご存じでしょうか。GPAI（Global Partnership on AI）とは、G7の議決のもと、AI活用が広がるにつれて健在化してきた人や社会への不利益といった問題を、哲学、政治、法律、産業、人権、そして科学などの様々な領域を越えて議論する国際的な専門組織です。2020年に結成され、3年目となる今年は29か国を代表して日本が議長を務めています。

3年目の開始にあたり、総務省と経産省の主導で年次総会GPAIサミットが11月21、22日に開催されました。富士通は、この機会に「信頼されるAIの実現にむけた課題」について世界的な合意形成を目指しGPAIサイドイベントのセッションを主催しました。本記事では、その様子を共有し、標準的なAIリスク・ガバナンスの実現に向けた富士通の取り組みをお伝えします。

目次

「信頼されるAI」を社会に浸透させていく上での課題

AIが経済成長への大きな期待を担う一方、カメラ映像による監視、人材採用、融資審査などにおいてAIによる判断が下されるなど、人権や社会価値へのインパクトも懸念されています。このような新技術を正しく活用するための知恵や知識が求められており、現在では多くの国がAI利活用の原則やガイドラインを策定しています。哲学を中心としたAI倫理の議論は成熟したと言えそうです。

これを受けて欧州連合（EU）がAI法案（AI Act）を2021年4月に発表。2025年の施行も計画されています。この欧州AI法案に違反した企業は多額の罰金を含む制裁の対象となるのですが、多くの国や業界がこれを受け入れる姿勢を示しており、本法案に準拠するための具体的なプロセスや手段の整備が今後の急務となっています。しかし、このためのアクションを誰が主導し、どのように解を提供するのか曖昧な部分も大きいようです。

そこで富士通は、日本がGPAI議長国として年次総会GPAIサミットを開催する場で、"The ways to trustworthy AI in practice" のパネルセッションを企画し、有識者らとともに前記の問いに関する共通理解を掲げました。そして、これらを富士通の研究開発に対する期待と捉え、これに沿うことをコミットしました。

パネルセッション “The ways to trustworthy AI in practice” について

パネリストは次のとおりです。いずれの登壇者も信頼されるAIの標準、方式、実装、技術の各面において際立った活躍をなさっている方々です。

パネルセッションの様子と登壇者

- Martha Russell博士、スタンフォード大学（ファシリテータ、写真上段中央）
  社会学と情報メディア科学の融合研究拠点mediaXをリード
- 杉村領一博士、産業技術総合研究所（写真上段右）
  ISO/IEC JTC1 SC42国内委員会 委員長として、AI倫理に関する標準規格をリード
- Matthias Holweg教授、オックスフォード大学（写真下段左）
  欧州AI法準拠のためのAI監査方式capAIを提唱
- Mikael Munk氏、2021.AI創業者CEO（写真上段左）
  AIガバナンスプラットフォームGraceを搭載する商用MLOpsのベンダー
- Ramya Srinivasan博士、米国富士通研究所（写真下段右）
  生成アートAIにおけるバイアスなど、AI倫理に関わる研究

特にRussell博士は分野横断の情報メディア研究に優れた経験をお持ちであり、富士通がAIの倫理について着手し始めた頃から、有識者や実践家らとの相互の学びの機会を設けるなどのご支援をいた

だいています。今回、ファシリテータを快諾いただき、欧州AI法案発表後に注目を集めるAI倫理の実践化の局面で、これに携わる人や組織がどのような枠組みのもと、どのような役割を担っていくべきかを問いかけるメッセージを共に事前検討しました。その結果、パネルセッション当日では次のような気づきを共有することができました。

パネルセッションでの気づき

- 政策立案、標準化、国際機関の各業界がそれぞれの共通フレームワークを構築することが課題
- 根拠に基づく政策立案や社会実装への転換には政策プロトタイピング（※1）や規制サンドボックス（※2）が有効
- 標準化や応用分野の専門家の意見が法規制に影響を与えていくことが健全
- 企業がAIをリスクのひとつと捉え、利害関係者との対話を経ることでAIリスク・ガバナンスを達成

※1　プロトタイピング：基本的かつ主要な機能を備えたモデル（プロトタイプ）を作成して事前検証を行い、施行（デプロイ）した場合の効果や影響を評価する手法

※2　サンドボックス：大規模な問題に発展してしまうことがないようコントロールされた環境下で、特例的に罰則等の適用を免除し、イノベーションを奨励する社会実証を行う場やその制度のこと

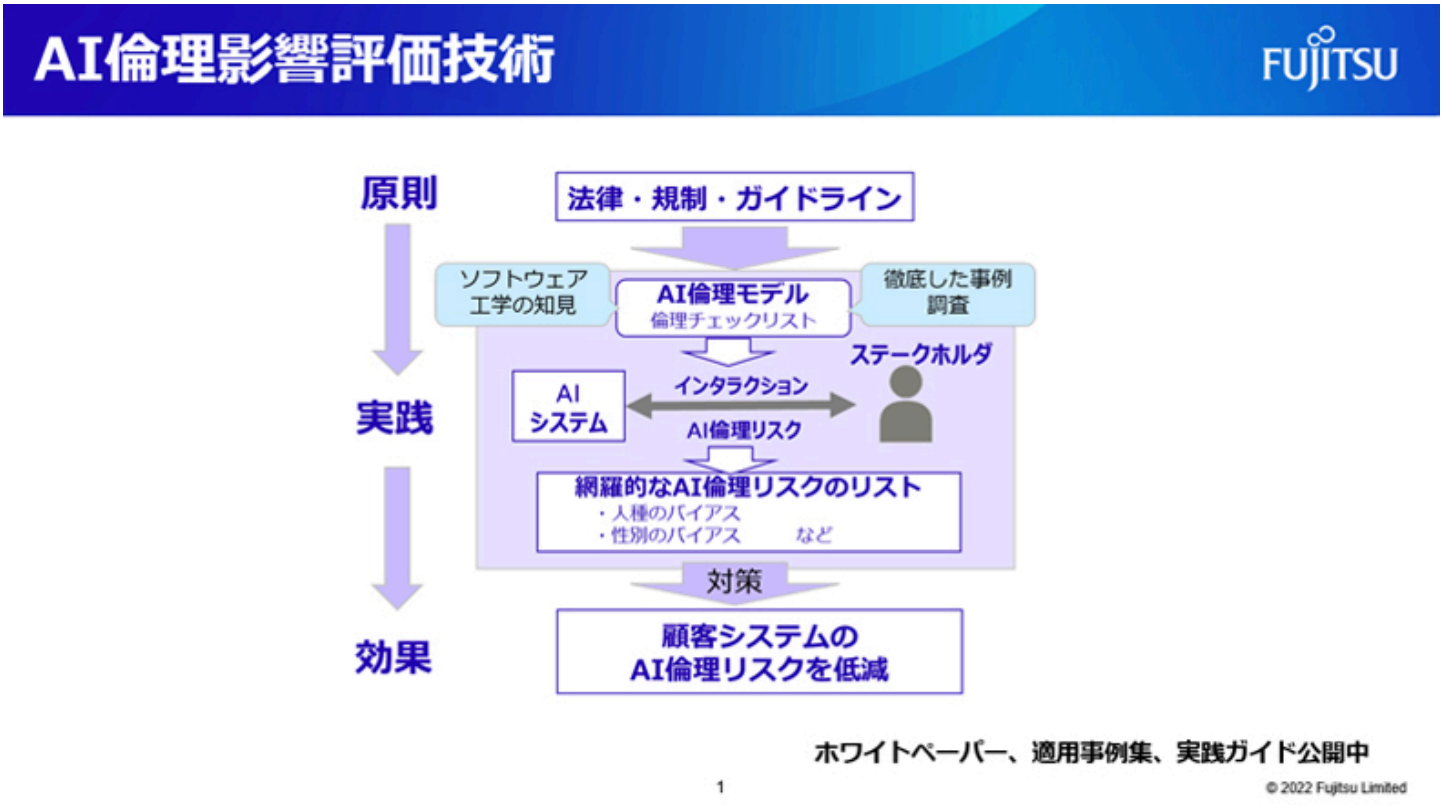
「AI倫理影響評価技術」 富士通が考える信頼できるAIとは

次に “The ways to trustworthy AI in practice” の中で、米国富士通研究所のRamyaが発表した、富士通が考える信頼できるAIの課題とその解決に向けた「AI倫理影響評価技術」をご紹介します。

富士通はAIを法や規範に準拠させ、それを証拠づけるためには手順を決めてそれに倣うのみでは不十分だと考えています。例えば、AI開発の各段階で実施すべき事項リストをチェックするような場合などが該当します。これにも増して重要なのが、法や規範を逸脱するリスクを査定する手段です。ここで言うリスクとは、AIを使って悪いことを考える人が出るかもしれない、意図せずに社会に悪を成してしまうかもしれない、そうしたことに気付かないかもしれない、などの可能性に該当する具体的な事象のことです。

AI倫理影響評価技術（図）を使うと、これから開発するAIが引き起こす可能性のあるリスクを査定できるようになります。同技術では、人、グループ、データ、アルゴリズムなどの間の相互運用の正しい形（AI倫理原則やそのための要件）と、これに反する形（過去にAIが引き起したインシデント）

を体系化した知識ベースを提供しています。これを参照することで実施されたリスク査定の結果は、追跡、検証、再現できるという裏付けのある法準拠の証拠を提供します。



AI倫理影響評価技術（AIシステムに潜むAI倫理リスクを網羅的に特定する技術）

今後にむけて

年次総会GPAIサミット開催後は産官学からの参加者から多くのフィードバックがあり、今回の企画そのものがグローバルな情報発信やパートナー共創への挑戦として優れていることへの評価や、具体的なトピックに関して意見を聞きたいといった専門性に対する評価を頂くことができました。富士通は、欧州におけるAI倫理原則の議論の場であるAI4Peopleに創設時から参画し、特に銀行・金融分野で貢献してきました。今回頂いたフィードバックは、欧州AI法案発表後のAI倫理の実践面においても、国際協調、学際的なアプローチが今後も富士通が目指すべき研究の進め方だと確信することができました。

またAIリスク・ガバナンスのための共通フレームワークの構築が重要というメッセージをお伝えすることができました。日欧米、英国、シンガポールなどの主導により、AIリスクへの対処手段に対して具体的な方式の策定や合意形成が今後進められていくと予想されます。私たちは、AI倫理影響評価技術がこれに役立つ強力なツールとして認知・共感してもらえるよう、業界全体と様々な地域とともに技術の強化と普及に注力していきます。

この記事を書いたのは

富士通研究所AI倫理研究センター プロジェクトディレクター

稲越 宏弥

博士（情報科学）。2016年よりディープラーニング・機械学習の研究に従事、2018年よりAI倫理原則の議論で先行する欧州（ロンドン）に駐在しAIとその倫理の研究を主導。欧州での経験で学んだ多様性を尊重する価値観を研究開発に活かす。