「AI 安全性とガバナンスをめぐる国際的な潮流」開催レポート

2024 年 3 月 28 日、東京大学未来ビジョン研究センター及び東京大学国際高等研究所東京カレッジは、「AI 安全性とガバナンスをめぐる国際的な潮流」と題したセミナーを開催した。本セミナーは、東京大学本郷キャンパス国際学術総合研究棟 SMBC アカデミアホール及びオンラインのハイブリッドにより開催され、オンラインでは 178 名、会場には約 30 名が参集した。

本イベントは、生成 AI の普及に伴い、AI の安全性(セーフティ)の議論が国内外で展開されている一方で、「安全性」の議論には様々な種類と対応策があり、さらに各国が置かれている状況や文脈により、何を「安全」とみなすか、あるいはどのような脅威・リスクが重視されるかは異なるという状況を踏まえ、また、AI Safety Institute が英国・米国・日本等で設立されている中、日本固有の「安全性」の種類とその対応策を把握しておくことが、今後の国際的な連携の基盤としても重要になるとの問題意識から開催されたものである。本イベントにおいては、海外の AI ガバナンスに関する専門家の出席を得て、国際的な AI 安全性やガバナンスの潮流に関する議論が交わされた。

登壇者

Merve Hickok: President and Research Director at Center for AI & Digital Policy (CAIDP)
Cyrus Hodes: Lead, SAFE project at the Global Partnership on AI (GPAI)
Inma Martinez: Chair of the Multi-stakeholder Experts Group, Global Partnership on AI (GPAI)
Michael Sellitto: Head of Global Affairs at Anthropic
飯田陽一: 総務省国際戦略局情報通信国際戦略特別交渉官
城山英明: 東京大学未来ビジョン研究センター教授
江間有沙: 東京大学国際高等研究所東京カレッジ准教授（司会）

（1）開会挨拶

最初に、東京大学未来ビジョン研究センターの城山英明教授から、開会挨拶があった。城山教授は、同センターの技術ガバナンス研究ユニットが、新興技術（Emerging Technologies）のガバナンスに焦点を当て、リスクチェーンモデルに関する研究や GPAI への参画を通じた AI の安全性に関する国際

的な議論に貢献してきた経緯を紹介しつつ、生成AIの爆発的な普及やG7広島AIプロセスといった近時の急速な情勢の変化に鑑み、AIの安全性に係る課題の整理や体制の整備についてそれぞれの現場に即した議論が求められている現状を踏まえ、本イベントが日本というコンテクストに係る議論のきっかけになることへの期待を示した。

（２）論点紹介

まず、パネリストからの論点紹介として、GPAIのInma Martinez氏から、GPAIが「すべての人のためのAI」として、公平性や脆弱な人々を包摂することを重視していること、また、これらを含めたGPAIの議論において日本が果たしてきたリーダーシップについて言及があった。加えて、直近のGPAIの活動が、G7広島AIプロセスの特色である共通の価値の実現に向けた合意形成を重視して行われているとの紹介があった。

また、Martinez氏からは、AIは、自動化にとどまらず、あらゆる産業分野に影響を与え変革をもたらされるものであり、そのようなAIの「安全性」は世界各国で文化的に多様な解釈がなされ得る一方で、「信頼できる」ということについては「技術的に機能する」とのコンセンサスがあるとの説明があった。その上で、GPAIはコンセンサスの形成を追求しているものの、AIに関する定義はもはや無効であって、それぞれの国の文化・価値に沿うべきであり、モノカルチャー化すべきではないことが強調された。

続いて、同じくGPAIのCyrus Hodes氏から、GPAIがマルチステークホルダーと連携して、生成AIの安全性の保証に取り組んでいる旨が述べられた。

その上で、Hodes氏は、生成AIのリスクとして、AIシステムの高度化に伴い、監査や評価の基準相互の不調整（Misalignment）が生じてくることが挙げられるが、その調整のためのインフラストラクチャーを構築する必要があるところ、AIセーフティー・インスティテュートとの協調への期待を示した。また、このほかに日本に期待することとして、AIの安全性に関する議論に係るマッピングへの協力及び国際的な議論への貢献を挙げた。

続いて、CAIDPのMerve Hickok氏が発言し、まず、同センターが、政府や国際機関への提言やAI政策に関するトレーニングの提供を任務としていることについて説明がなされた。次いで、アメリカのAI政策が政権を超えて一貫していること、政府機関向けの拘束力のある大統領令や、民間部門においても活用されうる任意のガイドラインなどの策定が進められているとの現状が紹介された。そして、ソーシャルメディアにおける有害情報を規制しな

かったという失敗への反省から、アメリカにおいてもAIセーフティー・インスティテュートが設立されことを解説しつつ、この種の組織をどの省が所管するかは当該国家が何に焦点を当てているかを示すものであるところ、アメリカでは欧州と異なり「安全性」の定義が広く、経済をも包含する概念であることも踏まえて商務省が所管することとされているとの見解が述べられた。そのほか、イギリスによるAIセーフティーサミット、韓国のミニバーチャルサミットといった最近の取組が紹介された。

その上で、Hickok氏は、AIと人権の議論においては、最低限の要素を共通するという意味での「相互運用性」が重要であり、広島AIプロセスに盛り込まれた諸要素に配意しつつ、マルチステークホルダーの参画を得て国際協調を進めていることが重要であると強調した。

最後に、AnthropicのMichael Sellitto氏からは、同社が責任あるスケーリングポリシー（Responsible Scaling Policy）の下で、バイオセーフティーレベルになぞらえたAI安全性のレベル（ASL）を設定し、リスクの程度に応じて安全性に係る対策を講じることとされていることが紹介された。その上で、昨年のAI開発のモラトリアムへの呼びかけにも言及しつつ、抽象的な危険のみに基づいて開発を中止するのではなく、ガバナンスの確保といった対策を講じることで対処するべきである旨が述べられた。

また、Sellitto氏は、広島AIプロセスにおいて策定された国際行動規範を非常に効果的な枠組であると評価しつつ、官民の協力によりコミットメントをモニタリングし、もって同規範への信頼を高めていくことへの期待を示した。

（３）パネルディスカッション

以上の論点紹介を受け、総務省国際戦略局の飯田情報通信国際戦略特別交渉官及び城山教授も加わり、江間准教授の司会により「AIガバナンスにおいて日本に期待することは何か」と題したパネルディスカッションが行われた。

まず、飯田氏から、海外事例の紹介を始めとする充実した発表への謝意と共に、各ステークホルダーによるAIの安全性に対する野心的な取組に対する賛辞が示された。また、飯田氏は、AI政策の多様性について、共通性や相互運用性を確保することの重要性を強調しつつ、各登壇者の発言からも、先進国間においてさえ、特にアプローチにおいてなお多様性があることを指摘した。また、Anthropic社の自発的な努力や国際連携への意欲について、このような取組に接し意を強くしたと評価した。

続いて、城山教授からの問題提起として、安全性とは何か、そして安全性はなぜ重要なのかという点が問いかけられた。次に、先端的なAIや生成AIがもたらす、従来のAIと異なる新たなリスクはどのようなものか、また、各国のAI政策を比較したときに、超党派の合意の有無や所管省庁が異なることは何を意味するのかという２つの点についてさらなる見解を求めた。

城山教授の問題提起に対して、まず、Sellitto氏からは、AIを巡る懸念やリスクは多岐にわたっているが、Anthropicが焦点を当てているところに即していえば、「安全性」とは、信頼できる安全な形でAIを使えるようにすることであるとの応答がなされた。

次に、Martinez氏は、21世紀は初めてあらゆる産業に安全がもたらされた世紀であると述べつつ、「安全性」とは、危害を加えず、危害を防ぐことであると指摘した。

これに対して、Merve氏は、AIの客観的な目的となる機能を出発点として信頼を考えることになるが、汎用目的技術としてのAIについては全てのユースケースを想定することができないことを指摘した。

また、Hodes氏は、汎用AIの時代では、あらゆるタスクがAIによる改善の対象となり得るが、そうした社会においてもAIシステムを調整することで価値観が維持される必要があると述べた。

これらを受け、飯田氏は、広島プロセスが生成AIのリスクを討議するために立ち上げられたが、後に基盤システムや先端的なAIも対象に追加された経緯を説明した。また、国際的な討議の中では、「安全性」と「信頼」が同時に議論され、安全性の定義に関する議論が避けられてきた側面があり、今後具体的な対策を講じる中で詳細な定義が必要との認識を示した。

司会である江間准教授からも安全性に関する議論においては、AI自体の安全性のみならず、法執行機関における活用などAIによって実現される安全性や、ほかの価値とのトレードオフの関係を含めて議論を枠付ける必要があるとの論点が提起された。

これに対して、飯田氏は、城山教授及び江間准教授の問題的はいずれも極めて重要なものであるとしつつ、技術ベースでイノベーションを進展させつつもリスクを最小限に留めるという点においては、政治・行政の各アクター相互で見解のギャップはそう大きくないとの見解を示した。また、飯田氏は、AIに係る政策立案プロセスにおけるマルチステークホルダーアプローチの重要性を改めて強調した。

Hodes氏は、飯田氏に賛同しつつ、二大巨頭としての米中という構図を指摘しつつ、日本のAIセーフティー・インスティテュートの設立といった取

組を賞賛し、調整役としての役割を果たしていくことへの期待を示した。

Merve 氏は、省庁による権限の違いに言及しつつ、それゆえにマルチステークホルダーの考え方が重要であることを強調した。

Martinez 氏は、欧州でもインターネットに係る規制の整備が遅々として進まなかった経緯に触れつつ、AI に関する規制が、日本の提言も踏まえて原則・価値・共通点を踏まえたグローバルの合意形成の下で進められたと述べた。

Sellitto 氏は、技術開発の初期段階では、規制がイノベーションを妨げるとの懸念が持たれうるが、徐々に何を規制すべきかが分かってくるものであるとしつつ、Anthropic の ASL も、まずは規制の策定及び実施を行い、そこから得られた教訓を公表するという実践であったと振り返りつつ、将来的には政府が同様の規制を導入することを期待していると結んだ。

（４）質疑応答

オンライン参加者からの、日本が近年サイバー攻撃の標的となっていることを踏まえ、AI の安全性や信頼性の担保に何が必要かという質問に対し、Sellitto 氏は、現在 AI のサイバーセキュリティに関する明確な指針はないものの、開発者向けのサイバーセキュリティースタンダートの形成が進められていることを説明した。また、Martinez 氏は、AI を標的としたサイバー攻撃は数多く、むしろこれから学ぶことでレジリエンスを高めることができるのではないかとの展望を示した。

（５）総括・閉会挨拶

本イベントの締めくくりに当たって、城山教授は、ここまでの議論及び質疑応答を総括して、「安全性」についてはあえて細かい定義を設けないのがよいように思われるが、共通の語彙やノウハウを整理していく必要性があると指摘した。また、AI への規制については、ハードロー・ソフトローという二分論は単純にすぎ、まずは抽象的な原則と経験の共有から学習のプロセスを進める必要があると提言した。

江間准教授は、参加者への謝意と共に、技術革新が急速に進展する中で、AI のセキュリティ及びセーフティ、ひいては AI ガバナンスについては、アジャイルかつ敏捷なプロセスを堅持する必要があるとの認識を示した。

最後に、東京大学国際高等研究所東京カレッジの星岳雄副カレッジ長が閉会の挨拶を述べた。星副カレッジ長は、本日の討議の重要性は明白であり、東京カレッジが未来ビジョン研究センターとともに本イベントを主催できた

ことは光栄であるとし、星副カレッジ長自身の専門分野である金融規制に関する議論を引き、常に金融システムの健全化を図るための努力が続けられているが、金融システムを完全に安全にできる規制やメカニズムは存在しないのと同様に、技術の安全性を確保するためには、その使い方にポイントがあり、危機を予防することと共に、危機を想定して対応策を備えておくことが必要であるとの所感が述べられた。また、星副カレッジ長は、AI のリスクを管理しつつ、人間中心の AI 開発を進めていく必要性に言及しつつ、本日の討議を今後の議論の皮切りにしたいと締めくくり、本イベントは盛会のうちに閉会した。