第８部

特集8 IAB Workshop on AI Control報告

浅井 大史

第１章　はじめに

2024年9月19日～ 20日の2日間にわたりIETFのIAB（The Internet Architecture Board）主催のワークショップであるIAB Workshop on AI Control[21]が開催された。本稿では、本ワークショップとその後の経過について報告する。なお、本ワークショップはチャタムハウスルール（Chatham House Rule）で開催されたため、参加者間で合意された報告書[22]に基づき報告する。

第２章　本ワークショップ開催の背景

まず、本ワークショップの背景について説明する。本ワークショップは、前述のとおり、IAB主催のワークショップである。IABは、インターネットの関係する社会課題に対する戦略や将来のインターネットアーキテクチャを検討にあたり、IETF内外のコミュニティやステークホルダーとの議論をする場としてワークショップを不定期で開催している。本ワークショップは、2023年頃から急速に利用が拡大し、注目を浴び始めた大規模言語モデル（LLM:Large Language Model）等の生成AIの普及に対して、これらのAIの学習において用いられる大多数のデータがインターネット上のデータであることから、AI学習のためのデータアクセス（とその制御）のあり方について議論するために開催されたものである。

データアクセスの制御機構としては、AI学習データの収集に利用されるCrawlerの制御方法としては、robots.txt（Robots Exclusion Protocol（RFC 9309））によるものがRFCとして定義されている。しかし、robots.txtは検索エンジンのためのCrawlerと収集したデータの利用方法（indexingを行うか等）を制御する目的のものであり、AI学習のような利用方法については想定されていなかった。そのような技術的な制御手法を含めたAIとインターネットを取り巻く技術的課題について、IETFとしての取り組み方針を検討するために、IABのMark Nottingham氏とSuresh Krishnan氏がチェアとなり、本ワークショップが開催された。

第３章　本ワークショップにおける議論

本ワークショップでは、まず本ワークショップ開催にあたり参加者から募集したポジションペーパー [23]の口頭発表に基づき議論が行われた。ポジションペーパーの発表には、EUのAI Actのような各国・地域の法規制との関係に関する指摘もあり、アクセス制御技術だけでなく、その技術が必要とされるシーンに対するポリシーやガバナンスについても議論の対象となっていた。また、従来のrobots.txtにおいても議論の対象となる著作権（copyright）の課題があったが、特にAIにおいては、従来のindexingを行うためのcrawlingとは異なり、マルチメディアを含めたコンテンツ自体を取り扱うため、そのようなケースをカバーできるような仕組みの必要性が議論された。また、AI学習がコンテンツを対象とするため、paywall（課金後にアクセスできるコンテンツ）に対するシグナリングや制御についての議論も行われた。

ポジションペーパーの発表後に、論点の整理が行われた。論点は、主に以下の3点があった。

- 用語の定義および整理
- 学習時と推論（生成）時でのデータの取り扱い
- 制御技術とその実効力

まず、1点目の用語の定義と整理については、技術標準化における議論を効率的に行い、目的からメカニズムを明確に説明するために必要不可欠なことである。一方で、AI技術が急速に発展する中で、ユースケースや適用ドメインごとに用語の定義が変わることがあり、本ワークショップ中に各用語の定義に関するコンセンサスを得ることはできなかった。簡潔な用語の定義の必要性は本ワークショップでも認識しているため、今後の議論においては随時定義が議論され、コンセンサスが取られていくものであると思われる。

2つ目の論点は、学習時と推論(生成)時でのデータの取り扱いについてである。生成AIは、大規模なデータセットからモデルを学習し、そのモデルを用いて利用者の入力(プロンプト)等に従ってコンテンツを生成するものである。この生成のプロセスは、機械学習・深層学習の用語を用いて、推論(Inference)と呼ばれる。生成AIの議論における複雑性のひとつは、学習と推論で
従来の機械学習・深層学習においては、学習データは最適化等に用いられ、学習データと同類・同様のデータが直接出力されることは希であった。しかし、生成AIでは、例えば、LLMは文章から学習をして文章を出力するため、学習データの内容と類似した文章が出力されることがある。また、画像や動画等のマルチメディアについても同様である。そのため、学習と推論が同時に議論されることが多い。一方で、Common Crawl[24]のように、データの取得がAIの学習目的に限定されるものではない(その他の科学技術の研究開発にも広く用いられている。)にもかかわらず、AI学習に広く用いられているという理由で、既存のrobots.txtを用いてCommon Crawlのcrawlingを拒否しているケースも報告された。このようなケースでは、crawling時ではなく学習時の利用目的に従ったシグナリング等の制御が必要であり、既存のrobots.txtでは実現できないことである。また、「商用利用禁止」といったようなコンテンツは、例えば、利益目的でない学術機関等による生成AIモデルの学習時においてはその条件に従ったものであると考えられるが、生成AIモデルは汎用なモデル(Foundation Model)であるため、そのモデルによる推論結果(出力)が商用利用禁止のコンテンツに基づくものであるのかを判別することは難しく、さらに、そのコンテンツ所有者側の意図もCrawling時、学習時、推論時で変化する可能性もある。このようなことも扱う仕組みについて検討する必要がある。

3つ目の論点は、制御技術とその実効力である。robots.txtのような手法以外に、HTTP headerでの制御やHTMLのmeta tagやJPEG等のメタデータ領域での制御などがあるが、HTML等はIETFでのスコープでないことからW3C等の多数の関連標準化団体との連携が必要となる。また、メタデータによる制御は、ファイルサイズの増大や誤った情報や意図しない情報の漏洩が起こりうるため、初期の検討では優先しない方針となった。一方で、制御技術が適切に処理され、その実効力がなければ、コンテンツ提供者は、広い範囲でのデータ利用を制限することになる。先述したCommon Crawlのcrawlingを拒否している事例のように、crawling自体を制限したり、botのアクセス自体を制限したりするようなことが起ってしまうと、インターネット・Web技術の発展自体を阻害するおそれがある。そのため、コンテンツ提供者の意図が正しく反映され、実効力を持つことも重要である。

上記の議論を受け、本ワークショップの結論としては、IETFにおいてWeb and Internet Transport (wit)AreaでWG化を目指す方針となった。

第4章　ワークショップ後および今後の展開

本ワークショップの議論を受け、IETF 121でBoFが開催された。その結果、ワーキンググループとして承認され、AI Preferences (aipref)WGとして活動を開始した。生成AI技術が著しい発展をみせる中で、日々変化する状況に対応しながらメーリングリストで活発に議論が行われている。IETF 122でもWGセッションが予定されており、また2025年4月にInterim meetingも予定されているため、今後も動向を注視しながら、WIDEプロジェクトの知見をもとにワーキンググループに貢献していく。