参考資料３

AI 学習に対する技術的な対応手段の一例について

クリエイターや著作権者等の中には、自らの作品を AI に学習されてほしくないという声や想いがある。
自らの著作物等が AI 学習に用いられることを望まない場合、技術的な対応手段によって、当該学習に用いられる可能性を低減することも可能と考えられる。
本資料は、一例として、技術的な対応手段のひとつである、ウェブサイト内の「robots.txt」の記載による方法について現況を事務局にて収集し、暫定的に整理、紹介するもの。

1.「robots.txt」の記載による収集制限の概要

○　AI 学習に用いられる学習用データは、インターネット上のウェブサイトに掲載されたデータを、自動収集プログラム（クローラ）を用いて広範に収集する手法によって収集されている例が多いと考えられる。

○　クローラによるデータ収集については、ウェブサイト内[1]の「robots.txt」という名称のファイルにウェブサイトの管理者が記載した制限（ロボット排除規約[2]）を尊重して収集を行う、という慣行が存在する。
「robots.txt」においては、特定のクローラについて収集を拒絶することや、収集可能とする部分（ディレクトリ）を指定するといった設定が可能とされている。

2. AI 学習のためのデータ収集と「robots.txt」への対応状況の一例

○　AI 学習に用いられる学習用データの収集においても、上記の「robots.txt」を尊重してクローラによる収集を行う旨を表明している例が多く見受けられる。

【参考】
《非営利団体「Common Crawl」の事例》
同団体はクローラによるデータ収集を行い、第三者に対してデータセットの提供を行っている（同団体の提供するデータセットは OpenAI 社の基盤モデル「GPT-3」の開発に用いられたとされるほか、非営利団体「LAION[3]」が提供する画像及びテキストのデータセット「LAION-5B」等の作成にも用いられたとされる）。

[1] ウェブサイトのルートディレクトリ（最上位階層）
[2] インターネット技術に関する標準化団体である IETF により”RFC 9309: Robots Exclusion Protocol“として標準化提案がされている。（https://datatracker.ietf.org/doc/rfc9309/）（令和 5 年 8 月 10 日最終閲覧）
[3] Large-scale Artificial Intelligence Open Network（https://laion.ai/）（令和 5 年 8 月 10 日最終閲覧）

同団体では、「robots.txt」ファイルに下記の内容を追加することで、同団体のクローラによる当該ウェブサイトの収集が停止する旨をウェブサイト上で表明している[4]。

```
User-agent: CCBot
Disallow: /
```

《OpenAI 社の事例》

同社は基盤モデル「GPT-4」等の開発や、対話型生成 AI サービス「ChatGPT」のサービス提供等を行っている。

同社では、「robots.txt」ファイルに以下の内容を追加することで、将来のモデル改善に使用する可能性があるデータを収集する同社のクローラ「GPTBot」による収集を禁止できる旨をウェブサイト上で表明している[5]。

```
User-agent: GPTBot
Disallow: /
```

また、同様に以下の内容を追加することで、ChatGPT のプラグインによるウェブサイトへのアクセスを禁止できる旨を表明している[6]。

```
User-agent: ChatGPT-User
Disallow: /
```

3. 著作権者等が取りうる技術的な対応手段の一例

○　著作権者等が、自らの著作物等が AI 学習に用いられることを望まない場合は、上記のような技術的手段による対応として、著作物等をインターネット上にアップロードするに際して、「robots.txt」において学習用データの収集に用いられるクローラによる収集を拒絶する（そのように設定された「robots.txt」を有するウェブサイトに著作物等をアップロードする）ことが考えられる。[7]

---

[4] 同団体 Web サイト内 ”How can I block this bot?” （https://commoncrawl.org/big-picture/frequently-asked-questions/）（令和 5 年 8 月 10 日最終閲覧）なお上記はウェブサイト全体の収集を拒絶する設定だが、いずれのクローラも、個別のディレクトリを指定して収集を拒絶することが可能。

[5] 同社 Web サイト内 ”GPTBot” （https://platform.openai.com/docs/gptbot）（令和 5 年 8 月 10 日最終閲覧）

[6] 同社 Web サイト内 “ChatGPT-User” （https://platform.openai.com/docs/plugins/bot）（令和 5 年 8 月 10 日最終閲覧）

[7] 当該著作物等が「robots.txt」の記載により収集が制限されていないウェブサイトに別途アップロードされている場合は、当該ウェブサイトから収集される場合があることや、ウェブサイト以外の情報源（書籍等）から収集される場合があることには留意する必要がある。