営業秘密官民フォーラムメールマガジン掲載コラム　第 38 回

AI・IoT の開発における学習用データセットの生成ノウハウの保護について

弁護士知財ネット
弁護士　　後　藤　　大

AI・IoT の開発における契約に関しては、経済産業省が、2018 年 6 月 15 日に「AI・データの利用に関する契約ガイドライン」（以下「契約ガイドライン」といいます。）を公表し、農林水産省においても、2018 年 12 月 26 日には、農業分野におけるデータ契約ガイドラインが公表され、2019 年 7 月 1 日から農業分野における AI の利用に関する契約ガイドラインの検討が始まっています。

本コラムの第 28 回（https://www.ipa.go.jp/security/economics/mailmag/20181017.html）においても、既に、契約ガイドラインと限定データの取扱いに関する不正競争防止法の改正が取り上げられています。

AI・データの利用に関する契約ガイドラインもこれから参照されることが増えていくと思われますが、現時点での AI の本質が、データ駆動型といわれるように、データから学習して予測を行う点にあることからすると、AI の開発においては何よりもデータの取扱いが重要になってきます。

今回は、学習用データセットに着目して、実務上悩ましいと思われる点について、話題を提供したいと思います。

契約ガイドラインにおいては、発注者であるユーザ側から提供されるデータについては契約書別紙に明細として記載されることを前提に「本データ」と定義し、これに対して、本データを AI 開発のために整形または加工したデータについては「学習用データセット」と定義しています。

具体的には、ユーザ側が画像データを提供し、これに対して、ベンダ側が AI による学習に適するよう学習用データセットとするべく、外れ値、欠損値の除去や、恣意的な値の除去、データ分布からの適切なデータの切り分け等や、どの部分が学習対象なのかを特定する等の加工を行います。

これは、おおよそユーザ側には、社内にデータサイエンティストがいるという企業が少ないことによるものです。

ユーザが提供した本データを学習用データセットに加工するという作業は、ベンダ側が行うことになりますが、ディープラーニングを含めた機械学習を利用した AI 開発においては、コモディティ化した類型の開発事案であれば格別、どのように学習用データセットを生成するかというのは、学習用プログラム、ハイパーパラメータの設定とともに、ベンダ側の試行錯誤の成果であり、その生成手法は開示したくないノウハウであります。

この点、契約ガイドラインで提案されているように、ユーザ側としては、ユーザ側が提供したデータを加工したものである以上、当該データの改変物はユーザに返還をする規程が盛り込まれていることになります。また、ユーザ側としては、加工されたデータを別のモデル生成に利用したり、別のベンダに改良を発注しようとすると、学習用データセットについては、手元に欲しいという動機付けが働きます。

しかしながら、上で述べたとおり、学習用データセットの生成は、ベンダ側の試行錯誤の成果であり、その生成手法はノウハウであるため、ベンダ側としてもこれを提供することはなんとしてでも避けたいところです。

そうすると、ベンダ側としては、AI 開発契約において、学習用データセットが、生データ、生データの改変物、秘密情報、秘密情報の複製（改変）物、又は中間成果物のいずれに該当しようとも、ユーザ側に開示しないで済むように、ユーザ側に提供することを除外する規定を契約上設ける必要があります。

とはいえ、ユーザ側との力関係等により、ベンダ側が、ユーザ側に学習用データセットを提供する場合もあり得ます。この場合、ユーザ側との間で、双務の秘密保持義務がある場合には、秘密情報であることを明記して開示するということになるでしょう。

もっとも、生データと学習用データセットを対比すれば、どのような加工をしているかはある程度わかってしまうため、ノウハウの流出は避けられないことになります。そうすると、そもそもベンダ側としては、秘密にしておきたいノウハウを利用したが句集用データセットについては、開示できないという選択肢しか残らないということになります。

他方で、ユーザ側は学習用データセットを入手すれば、自由に利用できるかというと、当然のことではありますが、秘密情報であることを明記された上で、学習用データセットの提供を受けたユーザ側は、当初の AI（「学習済みモデル A」とします。）の開発という利用目的の制限を受けることになりますので、これを別の AI（「学習済みモデル B」とします。）の開発に利用できるかというと、利用できないことになります。

そうすると、契約書上、ユーザ側は転用する意図があるのであれば、秘密情報からベンダ側から提供を受ける学習用データセットを除外してもらう必要があります。

もっとも、学習済みモデル B の生成に、新たにデータを追加した新たな学習用データセットを学習に利用すると、生成されるパラメータは、学習済みモデル A のものとは異なります。そして、パラメータからどのような学習用データセットが用いられたのかを解析することは困難であることからすると、紛争になったとして、ベンダ側としては、ユーザ側が目的外利用をしたことを証明することは、現在の技術では不可能かと思われます。

以上見てきたように、AI 開発における学習用データセットの取扱いひとつをとってみても、従来のシステム開発とは大きく異なり、ユーザ側ベンダ側双方が、長く協力して AI 開発及びその後の AI の精度の維持を図っていけるよう、AI 開発契約を締結する時点で、将来の運用も考えた条項を設定する必要があると思われます。

以上