特集《著作権》

生成 AI と依拠性についての調査研究

令和５年度著作権委員会第２部会

松田　真、井内　龍二、井上　正、前原　久美、前渋　正治、
松田　光代、辻村　和彦、高橋　信吾、田中　かおり、甲斐　一真

要　約

生成 AI を用いると、操作者が元の著作物を全く知らなかったとしても類似性の要件を満たした作品が生成され得る。その場合の依拠性をどう考えるか。類似性があれば依拠性が推定されてしまう実務でよいのだろうか？　令和５年度　著作権委員会　第２部会では、（ｉ）画像生成 AI を操作した結果生成された画像が、既存の著作物との間で類似性の要件を満たすこと。（ii）画像生成 AI を操作した人間が、類似する既存の著作物に依拠する意図がまったく無かったこと、若しくはその著作物の存在をそもそも知らなかったこと。（iii）生成に使用された画像生成 AI が、類似する既存の著作物を学習していること。の三つの条件を満たすことを仮定した上で、「福笑いモデル」と「拡散モデル」の２つのモデルを例示して議論を行った。本稿では、当該議論の結果を報告する。

目次

１．はじめに

本稿は、令和５年度の著作権委員会第２部会における議論をまとめたものである。

ChatGPT をはじめとして、生成 AI と呼ばれる人工知能のサービスが各種登場し、その利便性により急速に普及している。生成 AI については、その利便性が歓迎される反面、既存の情報を膨大に学習して成長するというその特性から、知的財産の問題について懸念が示されている。特に、画像生成の AI に関しては、画像生成 AI の成長によって人間のイラストレーターが不要になるとの懸念から、自ら手を動かしてイラストを生成するイラストレーターを中心として非常に活発な議論が行われている反面、感情的な言説も見られる。

新しい技術の隆盛により各種の問題が浮上することは生成 AI に限った特別な事象ではなく、時代の発展とともに古いものが淘汰され、若しくは不要になるという社会の原理とも言える事象であり、基本的には帰趨を見守る以外にはない。他方、人が人のために定めるルール、すなわち法に関しては、単に帰趨を見守るというわけにはいかず、新しい技術に社会としてどう向き合うべきかということを深く検討したうえで、社会に益する方向性を選択する必要がある。

即ち生成 AI について、社会に益となるものとして推奨するのか、社会に害となるものとして禁止するのかという二択であるが、実際にはその中間で社会にとって最も益となるバランスを見出すこととなる。

２．問題提起と前提条件

著作権の存在するイラストや写真の画像データを学習すること、学習のために AI の学習エンジンに入力することについては、根強い否定の声はあるものの、その著作権法上の扱いについては著作権法第 30 条の 4（以下、法律名の明記無き場合は著作権法）によって一応の決着を見たということになっており、本稿では対象としない。

本稿では、様々な著作物を学習した画像生成 AI を操作して画像を生成した結果、学習済みの著作物と類似性の要件を満たす画像が生成されてしまった場合についての、「依拠性」の考え方について論じる。

大前提として、画像生成 AI を操作するのは人間であり、人間自身に依拠の意図がある場合には、操作した画像生成 AI が類似性の要件を満たす著作物を学習していたか否かに関わらず依拠性は認められる。即ち、画像生成 AI によって生成された画像に類似する他の著作物が、画像生成に用いた画像生成 AI において学習されていなかったとしても、依拠性は認められるべきである。あくまでも画像生成 AI は人間が使用する道具であり、依拠したか否かは法的には人間自身の問題である。

他方、画像生成 AI を操作した結果生成された画像が、既存の著作物との間で類似性の要件を満たす場合において、画像生成 AI を操作した人間が上記の既存の著作物に依拠する意図がまったく無かった場合、特に、著作物の存在自体を知らなかった場合についてはどうだろうか。

このような状況において、その画像生成 AI が上記の既存の著作物を学習していた場合、直ちに依拠性が認められて然るべきだという言説がある。しかしながら、画像生成 AI、特に ChatGPT に代表される拡散モデルと呼ばれる種類の AI の仕組みを考えれば、そのように断定してよいものかどうかについて疑問が残る。そのような考えに基づいて依拠性を認定し、もって生成 AI を制限する方向にルールが定まってしまった場合、それによって守られるモノがあるとしても、その裏で社会に益する有益な技術の発展が阻害されることにもなる。

従って本稿においては、

（ⅰ）画像生成 AI を操作した結果生成された画像が、既存の著作物との間で類似性の要件を満たすこと。

（ⅱ）画像生成 AI を操作した人間が、類似する既存の著作物に依拠する意図がまったく無かったこと、若しくはその著作物の存在をそもそも知らなかったこと。

（ⅲ）生成に使用された画像生成 AI が、類似する既存の著作物を学習していること。

上記の 3 つを前提条件として、著作権侵害の判断の上で依拠性が認められるか否かについて論ずる。

３．生成 AI の例示

上記の通り、本稿の目的は「画像生成 AI によって既存の著作物との間で類似性を満たす画像が生成され、かつその画像生成 AI がその既存の著作物を学習していたとしても、AI の内部処理の仕組みによっては依拠性の要件が満たされない場合がある」という結論を導くことにある。そこで、まずは逆の場合、即ち、「学習されていれば依拠性あり」と断じられても仕方のない AI の内部処理の態様について具体的に例示する。

「学習されていれば依拠性あり」と主張する側による生成 AI の内部処理の認識は、多くの場合において本章で例示するような仕組み、例えば入力されたデータをそのまま保存して切り貼りするだけの、言うなれば低レベルな仕組みであると誤解しているのではないかと感じる。即ち、本章の例示及び次章の例示により、ChatGPT に代表されるような昨今普及している AI のモデルが、多く誤解されているような、学習のために入力されたデータを切り貼りするだけの低レベルなものではないことを明らかにする。

本稿における言葉の定義は、以下の通りである。

入力データ：生成 AI をトレーニングするために入力するデータ

課題データ：AI 学習のために与える課題を示すデータ

正解データ：課題データに基づいてたどり着くべき正解を示すデータ

内部データ：入力データを学習した結果として AI 内に保存されるデータ

命令データ：トレーニングされた生成 AI に生成を実行させるために与える命令（主に生成したいものを示す）データ

生成データ：トレーニングされた生成 AI に命令を与えた結果物として生成されるデータ

３．１　福笑いモデル

「入力されたデータをそのまま保存して切り貼りするだけ」というモデルである。そのようなモデルを直感的に示す例として「福笑いモデル」と名付ける。なお、3.1 および 3.2 において、完全なおたふくの画像が「既存の著作物」であると仮定して論ずる。

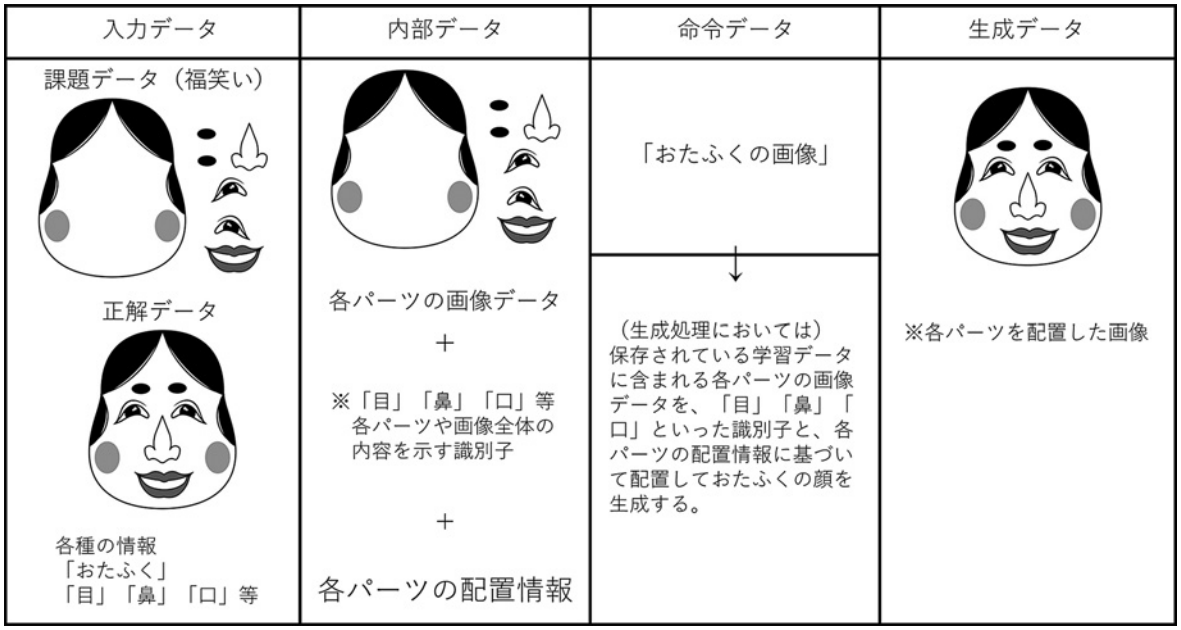

図１　福笑いモデル

福笑いモデルにおける入力データは、課題データとしての福笑い画像、即ち、「顔の輪郭」「目」「鼻」「眉」「口」それぞれの画像データと、正解データとしての完成されたおたふくの画像と、正解が「おたふく」であることや、各種のパーツがそれぞれ「目」「鼻」「眉」「口」であることを示す識別子である。このような入力データに基づく学習としては、課題データに含まれる各パーツの画像を配置し直して正解データの画像を再現するための学習が行われる。その結果として、各パーツの配置関係を示す情報や、「目」「鼻」「眉」「口」等の各パーツの内容を示す識別子が内部データとして保存されることとなる。

生成段階において、既存の著作物であるおたふくの画像に依拠する意図なく、一般的なイメージとしてのおたふくの画像を生成する意図で「おたふくの画像」という命令データが与えられると、福笑いモデル AI は「おたふくの画像」という命令に基づいて内部データに含まれる識別子を検索することにより、「おたふくの画像」という識別子に関連付けられている各パーツの画像データを抽出し、抽出した各パーツの画像データを配置情報に基づいて配置することで画像を生成する。

これにより、「おたふくの画像」という命令を満たす画像データが生成データとして生成される。

このような福笑いモデルによれば、「入力データ」に基づいて生成された「内部データ」は、既存の著作物であるおたふくの画像そのものではないにしろ、各パーツの画像に付された識別子や配置情報に基づいて完全なおたふくの画像を復元可能なデータである。従って、生成データとして入力データのおたふくの画像と全く同一の画像が生成される場合があり、その場合に生成される画像は、入力データの画像を一度パーツごとにバラバラにしたものを福笑いのように再構成したものという事になる。つまり、「入力データ」に依拠して「内部データ」を生成し、その「内部データ」に依拠して「生成データ」を生成しているのであり、それぞれの依拠のタイミングにおいてひとつ前のデータを復元可能であるため、類似性も認定されうる状態と言える。換言すれば、「入力データ」、「内部データ」、「生成データ」はデータとして連続している。このような場合、最終的に生成された「生成データ」は「入力データ」に依拠していると言えるのではないか。

そもそも、著作権法において画像に類似性があるという状態は、画像（イラスト）が著作物たりえる要因、すな

わち創作性を発揮する要素が類似している状態である。福笑いモデルにおいては、既存の著作物である「正解データ」が著作物として創作性を発揮する要素が「内部データ」にも残っている状態である。そのような「内部データ」に基づいて生成された「生成データ」が「正解データ」と類似しているということは、「内部データ」に残っている上記の要素が「生成データ」に用いられているという事である。これはまさに「複製」であり、データとして「依拠」していると言えるだろう。

３．２　拡散モデル

次に、昨今普及している“画像”生成 AI に主に用いられているモデルについて説明する。そのようなモデルは拡散モデル（Diffusion Moldel）と呼ばれる。尚、拡散モデルを詳細に説明すると、説明量が膨大かつ専門的になり過ぎてしまうため、本稿では入力データに対する依拠性の議論において最低限必要な程度に簡略化して説明する。

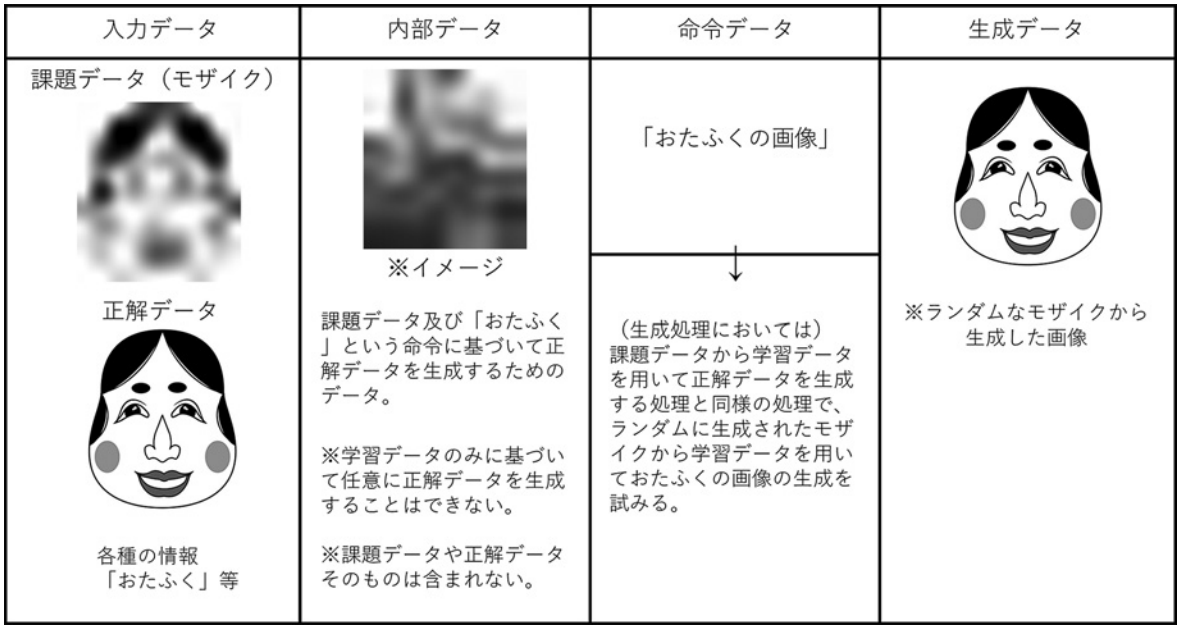

図２　拡散モデルの概要　その１　モザイクパターン

図２は、拡散モデルの一つのパターンとして、モザイク処理によるパターンを示す図である。尚、拡散モデルの理解を容易化するための例としてモザイク処理を例示するが、実際の拡散モデルにおいてはモザイク処理とは異なる処理（図３で示す処理）が用いられる。

図２に示す通り、モザイクパターンにおける入力データのうち正解データを完成されたおたふくの画像とすると、課題データとして用いられるのは、そのおたふくの画像の解像度を落としてモザイク状にした画像である。また、入力データには入力された画像の内容、すなわち「おたふく」であることを示す識別子を含む。このような入力データに基づく学習は、モザイク状である課題データに基づいて正解データを復元するための画像処理を想定し、その処理において正解データにたどり着くために必要であった各種のパラメータが逆算され、「おたふく」という識別子と共に内部データとして保存される。

生成段階において、既存の著作物であるおたふくの画像に依拠する意図なく、一般的なイメージとしてのおたふくの画像を生成する意図で「おたふくの画像」という命令データが与えられると、ランダムなモザイクパターンが生成され、そのモザイクパターンをベースとして、「おたふく」という識別子が付されたパラメータを用いて上記の画像処理が実行されることにより画像が生成され、生成データとなる。

図３は、拡散モデルのパターンであって図２とは異なり、より実際の拡散モデルに即したパターンであるノイズパターンを示す図である。図３に示す通り、ノイズパターンにおける入力データは、正解データとしての完成されたおたふくの画像と、課題データとしておたふくの画像にノイズを重畳した画像を含む。また、入力された画像の内容、すなわち「おたふく」であることを示す識別子を含む。

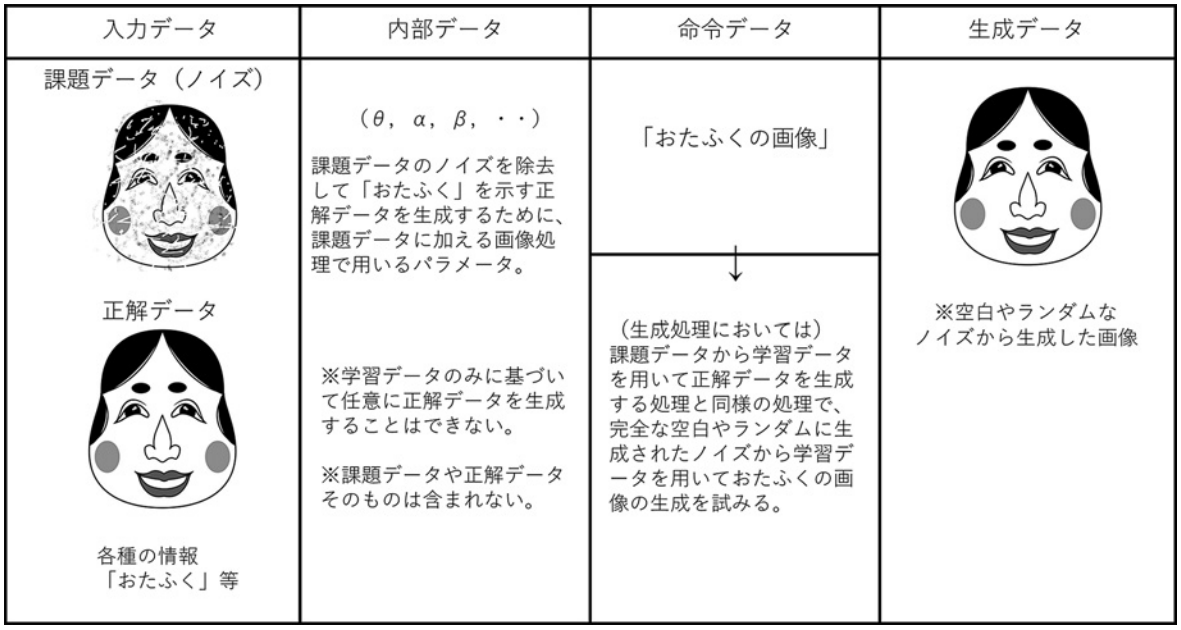

図３　拡散モデルの概要　その２　ノイズパターン

なお、図３の課題データにおいてはノイズが重畳されながらも元はおたふくであることがわかる画像を例示しているが、実際にはノイズの重畳を繰り返すことで純粋なノイズにまで拡散された状態である。このような入力データに基づく学習は、ノイズ状態である課題データに基づいて正解データを復元するための画像処理を想定し、その処理において正解データにたどり着くために必要であった各種のパラメータが逆算され、「おたふく」という識別子と共に内部データとして保存される。

生成段階において、既存の著作物であるおたふくの画像に依拠する意図なく、一般的なイメージとしてのおたふくの画像を生成する意図で「おたふくの画像」という命令データが与えられると、ランダムなノイズパターンが生成され、そのノイズパターンをベースとして、「おたふく」という識別子が付されたパラメータを用いて上記の画像処理が実行されることにより画像が生成され、生成データとなる。

図２、図３においては拡散モデルを最大限簡略化して説明しているが、実際には更に複雑な処理が繰り返し行われる他、生成 AI として実用性を有するまでには膨大な数の入力データについて学習を繰り返すこととなる。

このようなモデルによれば、「入力データ」に基づいて生成された「内部データ」は、モザイクからの復元処理やノイズの除去処理、ノイズパターンからの画像の復元処理等において実行される画像処理の細部を設定するためのパラメータであり、画像データではない。そして当然ながら、「正解データ」や「課題データ」は学習の処理においてのみ用いられ、内部データとしては保存されていない。

そのようなパラメータである「内部データ」を用いて「課題データ」に対して想定されている画像処理を実行することにより「正解データ」を復元することは可能であるが、「内部データ」のみに基づいて任意に「正解データ」を復元することは不可能である。

上記の通り様々な入力データに基づいて膨大な回数の学習を繰り返すことによってトレーニングされた生成 AI であれば、様々なパラメータが蓄積されることで、特定の正解データに近しい結果が出力される可能性が低くなる。即ち、様々な「入力データ」に基づく学習を繰り返すことで、特定の「入力データ」に類似する画像が生成される可能性は低くなる。

このように、拡散モデルにおいては、図１において説明した福笑いモデルのように入力データを構成する画像のパーツが保存されているわけではないので、そもそも入力データに類似する画像が生成される可能性は低いもののゼロではない。しかしながら、仮に入力データの画像に類似する画像が生成されたとしても、それは「入力データ」に依拠したとは言えず、様々な条件が重なることにより偶然生じたものである。なぜなら、上記の通り「内部データ」の段階において「入力データ」は完全に消失しており、かつそれを任意に復元することは不可能となっているからである。また、「内部データ」は上記の通り画像データではなく画像処理に用いられるパラメータであり、

情報として形式の異なるものであるため、「内部データ」と「正解データ」との間に類似性を観念することも不可能である。即ち、「入力データ」→「内部データ」→「生成データ」はデータとして連続していない。

また、図１において言及した著作物としての創作性の観点でも同様である。仮に、「正解データ」が著作物として創作性を発揮する要素データ A が「生成データ」に含まれていたとしても、上記の通り、「内部データ」にはそのような要素データ A は含まれず、少なくとも「入力データ」→「内部データ」→「生成データ」の間で創作性を発揮する要素のデータ A は連続していない。そして、上述した拡散モデルにおける画像生成の仕組みからすれば、「生成データ」に含まれるデータ A は、生成の際にランダムなモザイクやノイズパターンをベースとして「内部データ」に基づく画像処理を行うことで新たに「発生したもの」であると言える。

このように、拡散モデルにおいて学習に用いられた「正解データ」に類似する画像が「生成データ」として生成されたとしても、それは「正解データ」に依拠したものではなく、偶然に似てしまったものだと言えるのではないだろうか。

以上に対しては、人間の創作者においても、長年にわたって多数の既存表現物（上記の「入力データ」に相当）に触れて様々な刺激・影響を受けた結果として、依拠の意識のないままに、相当以前に一度触れた既存表現物 A に類似した表現物（上記の「生成データ」に相当）を作成するということはあり得、そのような場合に依拠性を否定するのは相当ではなく、そうであるとすれば拡散モデルもこれと変わらないのではないかとの指摘が想定される。

確かに、人間の創作者の場合には、その頭の中を覗いて解析することはできないため、創作者が A という「表現」に依拠したのか、あるいは多数の既存著作物に触れて刺激・影響を受けたことによって生じた「アイデア」に依拠したのかを区別することは不可能であるから、A に触れて A に類似した表現物を創作した以上は、事実認定における経験則上、依拠性ありとされてもやむを得ないところであろう。

しかしながら、生成 AI においては、内部でどのような処理がなされた結果として、A に類似した表現物（生成データ）が生成されたのかを解析することが一応可能である。上記の例で言えば、内部データとして各パーツの画像データを持つ福笑いモデルにおける生成データは、その内部処理として、各パーツの画像データという「表現」に依拠して生成されたものと評価できようが、一方で、内部データとして画像データを持たず、パラメータを持つにすぎない拡散モデルにおける生成データは、その内部処理として、画像データという「表現」に依拠して生成されたというよりも、画像処理のルールを設定するパラメータという「アイデア」に依拠して生成されたものと評価し得る場合もあるのではなかろうか。依拠性は、あくまで事実認定の問題であるところ、生成 AI については「表現」に依拠しているのか、あるいは「アイデア」に依拠しているのかを解析によって事実認定することが可能な場合もあり得るのであるから、頭の中の解析が不可能な人間の創作者の場合のやや概括的とも言える経験則が、生成 AI にそのまま妥当すると考えるのは安易にすぎるように思われる。生成 AI については、学習のための入力データの内容や内部処理の解析によって、依拠性に関連する事実を細かく認定することが可能である以上は、人間の創作者の場合のやや概括的な経験則に任せるのは相当でなく、生成 AI の内部処理に即したより精緻な事実認定を目指すのが、本来のあるべき姿というべきである。

４．小括

以上、生成 AI の例として、福笑いモデルと拡散モデルを簡単に説明したが、このような AI の詳細な仕組みを検討することなく一律に「学習されていれば依拠性あり」と断ずるのは早急ではないかというのが本稿の結論である。少なくとも、拡散モデルであれば「学習されていても依拠性なし」と言える場合が多分にあるというべきである。

５．補足、さらなる問題提起

もっとも、拡散モデルであれば常に「学習されていても依拠性なし」と言えるわけではなく、実際には個別具体的な検討が必要である。

注意すべき点の一つとしては、拡散モデルであっても学習が偏っていれば特定の著作物に類似した画像が生成さ

れる可能性が非常に高くなるということである。例えば、特定の作品や特定のクリエイターの作品を彷彿とさせる作品を生み出すことを目的として、特定の作品や特定のクリエイターに関する学習を繰り返すことで AI をトレーニングした場合、その AI が生成する画像は、特定の作品や特定のクリエイターの作品に類似したものとなる可能性が高くなる。

このようなモデルであっても、上記と同様に「内部データ」の段階で「入力データ」は残っておらず、データの連続性は途切れている。しかしながら、特定の作品に類似したものを生成するようにトレーニングされたモデルであり、結果的に生成される画像が特定の作品や特定のクリエイターの作品に類似したものばかりである場合には、上述したようにデータの連続性が無いことを根拠に生成段階で偶然類似してしまったものであるという論理を適用することには疑問を持たざるを得ない。

「拡散モデルだから」という短絡的な理由ではなく、それぞれにトレーニングされた生成 AI の特性に基づき、個別具体的に適切な判断がされるべきであろう。

上述した通り、本稿の前提は生成 AI を操作する人間に元の著作物に依拠する意図がなく、かつ当該著作物（上記の例における「正解データ」）のことを知らなかった場合、即ち、人間による意図的な依拠が完全に否定できる場合である。

これに対して、操作者に依拠する意図は無いものの、元の著作物を知っていた場合についてはどうであろうか。生成 AI を操作して生成された画像が、偶然にも操作者が知っている画像と類似していたような場合である。操作者が何らかの商業的な活動に用いるために画像を生成したとして、このように偶然にも知っている画像と類似している画像が生成されてしまった場合、操作者はその画像を意図的に排除しなければいけないのであろうか。

現在の実務においては、既存著作物との間で類似性が認められた上でかつ創作者が既存の著作物を知っていた場合や、既存の著作物にアクセスしていた可能性が高い場合には、依拠性が認定される。人間の創作者の場合には、実際の創作の際に、既存の著作物を意識していたかどうかを客観的に判定することが事実上不可能である以上、このような依拠性の認定実務には一定の妥当性があるといえよう。

一方で、AI 生成物の場合にはどうであろうか。例えば、操作者が生成 AI を 1 回のみ操作して生成された画像が偶然にも既存の著作物と類似していた場合、操作者の主観としては既存の著作物に依拠したという意識は皆無であろう。そのような場合にも、偶々操作者が既存の著作物を認識していた場合には依拠性が認定され、生成された画像を使用することができないという結論には、仕方のない部分があるとしても、なお疑問は残る。依拠性の判断は、作成した表現物が既存の著作物に類似していることを前提に、既存著作物に依拠することなく偶然に類似することがあり得るのかという観点からも検討されるべきところ、かかる観点からは、生成 AI による生成データについては、単に操作者の知不知のみを基準に考えるよりも、例えば、プロンプトの内容や多数回出力した場合に既存著作物に類似した生成データが出てくる頻度（偶然性）などの事情をも考慮して判断するのが相当ではないか。

なお、上記のような「操作者が知っていた場合には排除しなければいけないのか」という疑問を想定すれば、その延長上には本稿の議論が無意味に帰するような問題にたどり着く。即ち、操作者に既存の著作物に依拠する意思がなく、かつ操作者が実際にはその著作物を知らなかった場合であっても、その著作物がある程度の知名度を有しており、かつインターネット等で公開されることにより誰でもアクセス可能な状態におかれていたような場合には、操作者がその著作物を知っていたと認定される可能性は否定できない。このような場合の現状の訴訟実務における依拠性判断は、乱暴にまとめれば「類似性あれば依拠性あり」とも言える状態に見える。インターネットにより、様々な情報にアクセスすることが容易になった結果として「知らなかったわけがない」という前提があるのではないだろうか。

上述した議論によって生成 AI 自体に依拠性がなく、その操作者も既存の著作物を知らず、当然依拠する意図が無かった場合であっても、仮に「類似性あれば依拠性あり」「知らなかったわけがない」という実務が横行して著作権侵害が認定されてしまうという懸念が正しいとすれば、本稿の前提条件が成立する場面は極めて限定されることとなり、本稿における議論は完全に無意味となる。

人間の頭の中を覗くことが不可能である以上、依拠性を真に判断することには限界があるが、本稿において示し

た生成 AI と依拠性に関する議論をきっかけとして、ワン・レイニーナイト・イン・トーキョー事件において示された「偶然の暗合」は著作権侵害にならない、との原則に今一度立ち返った議論が行われることを期待する。

以上

（原稿受領 2024.8.20）