



생성형 AI의 학습데이터 공개 관련 논의와 입법 과제

-특정 화풍의 이미지 생성 사안을 중심으로-

백지연

최근 특정 화풍으로 사진을 변환하는 생성형 AI가 유행하면서, 학습 과정에서의 저작권 침해 여부를 판단하기 위한 학습데이터 공개가 쟁점으로 부상했다. 그러나 현재 관련 의무 규정은 없으며, 이를 두고 저작권자와 사업자 간의 입장 차가 크다. 미국은 관련 법안을 추진 중이고, 유럽연합은 「인공지능법」에 학습데이터 공개 규정을 도입했다. 학습데이터 공개 의무 규정의 단계별 입법이 이루어질 필요가 있으며, 「인공지능 기본법」과 「저작권법」에서 해결해야 할 과제를 구분하여 검토해야 한다.

1 학습데이터 공개 관련 논의의 배경

최근 SNS 이용자들 사이에 ‘지브리’ 화풍으로 이미지를 생성하는 유행¹⁾이 확산하고 있다. 그 결과 3월 말 특정 화풍의 이미지 생성이 가능한 모델이 탑재된 Chat GPT는 월간 사용자가 전달 대비 122만 명 증가했다.²⁾

2024년 12월 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 “인공지능 기본법”)이 통과되었고 2026년 1월 시행될 예정이다. 법안 심사 당시 문화체육관광부(이하 “문체부”)는 인공지능 투명성 확보 의무 규정³⁾을 신설하는 수정의견을 제시하였으나, 제정법을 통과시키고 부족한 부분은 보완하는 것으로 결론이 나며 최종안에 반영

되지 못하였다.

그 후 문체부는 2025년 업무보고 당시 질의응답에서 「인공지능 기본법」 및 「저작권법」 개정안을 준비한다고 언급했고⁴⁾, 최근 3월 ‘문화한국 2035’를 발표하며 “생성형 AI의 데이터 학습·활용 과정에서 저작권 등을 침해할 우려가 있으나 선진적인 법체계는 부족한 실정”이라고 밝힌 바 있다.⁵⁾ 이와 관련해 하위법령을 마련 중인 과학기술정보통신부는 최근 「인공지능 기본법」이 진흥법이라고 밝히며, “인공지능기본법 시행령에 저작권 침해 여부를 포함하기는 어렵다”라고 덧붙였다.⁶⁾

이 글에서는 특정 화풍의 이미지 생성으로 촉발된 학습데이터 공개 관련 논의, 이를 둘러싼 쟁점 및 해외 입법 동향을 정리하며, 향후 입법 과제를 제시한다.

1) 이용자가 Chat GPT에 “지브리 화풍으로 만들어줘”라고 입력하면 이용자가 업로드한 사진을 일본의 지브리 애니메이션 스튜디오의 화풍을 적용해 새로운 이미지를 생성해준다.

2) 아이지에이웍스 모바일인덱스 「3월 월간 인기 모바일 앱순위 총정리」(최종검색일: 2025.4.18.), <https://www.igaworksblog.com/post/march_chart>

3) 제31조(인공지능 투명성 확보 의무) 제4항: “인공지능사업자는 생성형 인공지능의 개발·활용에 사용된 학습데이터의 목록을 공개하여야 한다”

4) 「2025년도 문체부 업무보고」, 『대한민국 정책브리핑』, 2025.1.10.

5) 문화체육관광부 보도자료, 「문화한국 2035 발표」, 2025.3.6.

6) 나선훈, 「과기정통부 “AI 기본법은 진흥법”...저작권 침해 논의는 문체부로」, 『아주경제』, 2025.4.8.(최종검색일: 2025.4.28.), <<https://www.ajunews.com/view/20250408160740549>>

2 특정 화풍 이미지 생성과 학습데이터

최근 논란이 된 특정 화풍의 이미지를 생성하는 행위는 두 가지 측면에서 논의될 수 있다. 첫째 생성형 인공지능(이하 ‘AI’)을 통해 생성된 결과물이 다른 저작물의 저작권을 침해하였는가와, 둘째 생성형 AI의 학습 과정에서 다른 저작물의 저작권을 침해하였는가이다. 두 번째 침해 판단에서 학습데이터의 공개 여부가 중요한 역할을 한다.

「저작권법」은 아이디어가 아닌 ‘표현’만을 보호한다. 특정 화풍은 구체적인 표현이 아닌 아이디어의 영역에 속한다. 단순히 특정 화풍을 사용해서 채색, 선 등의 변화를 준 이미지는 기존 저작물의 표현을 그대로 사용하는 것이 아니므로 이렇게 생성된 결과물이 다른 저작물의 저작권을 침해하였다고 보기 어렵다.⁷⁾

남은 문제는 생성형 AI의 학습 과정에서 다른 저작물의 저작권을 침해하였는지에 관한 판단이다. 저작권자의 동의 없이 생성형 AI가 특정 화풍으로 대표되는 저작물을 학습하여 유사한 화풍의 이미지를 생성하였다면 「저작권법」에 저촉될 수 있다. AI 모델을 학습시키기 위해서는 저작물 또는 데이터를 훈련하고자 하는 주체의 서버에 가져오는 것이 필요하다. 웹에 존재하는 데이터를 전송받아 서버에 저장하는 과정에서 저작물의 전송과 복제가 이루어진다.⁸⁾ 학습데이터를 이용하기 위해서는 원칙적으로 데이터세트에 포함된 개별 저작물의 저작자로부터 이용허락을 받아야 한다.

현재 출시된 AI 모델들은 웹에 존재하는 데이터

를 많이 활용하는 것으로 보이는데 인터넷에 공개된 웹사이트에는 개별 저작물의 이용허락에 관한 내용이 포함되어 있지 않은 경우가 많으며, 인터넷에 공개했다는 사정만으로 저작물에 대한 이용허락이 이루어졌다고 볼 수 없다.⁹⁾ 결국 생성형 AI가 학습한 데이터들에 저작권 제한 사유(공정이용 등)가 적용되지 않는다면, 이는 적법한 권원이 없는 저작물의 무단 이용에 해당한다. 다만 생성형 AI가 특정 화풍의 이미지를 생성했다고 해서 특정 이미지를 학습했다고 단정 지을 수는 없다.¹⁰⁾

문제는 현재로서는 생성형 AI가 어떤 데이터를 학습했는지 공개되고 있지 않아 정당한 권원에 의해 저작물이 이용되었는지를 판단할 수 없다는 점이다. AI 개발자·사업자(이하 “사업자”)가 생성형 AI에 학습된 데이터를 공개할 의무가 없기 때문이다.

전 세계적으로 이미 학습 과정에서의 저작물 무단 이용에 관해 여러 저작권 침해 소송이 진행되고 있다.¹¹⁾ 많은 소송에서 원고는 자신들의 저작물이 무단으로 이용되었다고 주장하고, 사업자는 그렇지 않다고 반박하며 대립하고 있다.

3 학습데이터 공개 의무화 관련 쟁점

학습데이터 공개 의무화와 관련하여 권리자(저작권자)와 사업자는 상반된 의견을 보인다. 공개를 찬성하는 측의 주장을 살펴보면, 첫째 공개된 학습데이터 목록은 저작권 침해 판단에서 의거성¹²⁾ 증

7) 참고로 일본 문화청은 ‘화풍’에 관한 더 구체적인 답변을 내놓은 바 있다. 2024년 3월 발표한 보고서 「인공지능과 저작권에 대한 고찰」에 따르면, “작품·화풍 같은 아이디어가 유사할 뿐 기존 저작물과의 직접적인 유사성이 인정되지 않는 생성물은 저작권 침해가 아니다”라고 명시했다. 한국저작권위원회, 「일본 문화청, AI와 저작권에 관한 개념 초안 발표」, 『이슈트렌드』 제23호, 2024, p.1.

8) 이대희, 「인공지능 저작권 쟁점 분류」, AI-저작권 제도개선 워킹그룹 발제 자료, 문화체육관광부·한국저작권위원회, 2023, p.59.

9) 이철남, 「생성 AI의 데이터 학습에 관한 쟁점과 TDM 관련 저작권법 개정 방향」, 『과학기술법연구』 제29권 제3호, 2023, pp.46~49.

10) AI는 대량의 데이터를 기계 학습하여 알고리즘에 반영하는 방식이기 때문에 학습하지 않은 내용에 대해서는 결과물을 생성하지 못하게 되지만, 생성형 AI의 경우는 실제 사실에 기반하지 않더라도 언어 시각적으로 연계성이 높은 결과들을 조합하여 결과를 제시할 수 있는 것으로 알려졌다. 정준화, 「인공지능의 내재적 위험과 입법·정책 과제 -데이터·기술·이용자를 중심으로-」, 『NARS 입법·정책』 제162호, 국회의원조사처, 2024, p.6.

11) 대표적으로 ‘Getty Images (US), Inc. v. Stability AI, Inc.’, ‘Authors Guild, et al. v. OpenAI, Inc.’ 등이 있다.

12) 의거성이란 기존 저작물을 인식하고 이에 근거하여 만들었는지를 말한다.

명 자료로 활용될 수 있다. 현재는 AI가 어떤 학습 데이터를 토대로 산출물을 생성했는지 외부에서 알 수 없어 의거성을 증명하기 어렵다. 따라서 주관적 요건인 의거성은 침해자가 권리자의 저작물에 접근할 수 있었다는 사실과 두 저작물 사이에 유사성이 있다는 간접사실로 추정될 뿐이다.¹³⁾ 특정 저작물이 학습되었음이 공개되면 저작권 침해 판단에서 의거성이 입증될 수 있다.¹⁴⁾

둘째, 학습데이터의 공개를 통해 이용된 저작물이 확인되면 이는 저작권 이용허락 계약 체결의 기반이 될 수 있다. 생성형 AI는 다량의 학습데이터가 필요하며, 데이터는 생성하거나 구매하는 방식으로 확보된다.¹⁵⁾ 사업자들은 데이터를 구매하여 법적 위험을 줄이고 있으며, 실제로 콘텐츠 제공자와 이용계약을 체결하는 사례가 늘고 있다.¹⁶⁾

셋째, 학습데이터의 공개는 생성형 AI의 환각 현상과 편향성을 예방하고 신뢰도를 높인다. 데이터 공개가 불투명하면, 사용자는 AI 생성물의 품질을 신뢰하기 어렵다. 예를 들어, 특정 저작물을 학습하지 않고서도 AI가 해당 화풍의 이미지를 흉내 낼 수는 있겠지만, 그 품질은 실제 학습한 AI보다 낮을 수밖에 없다.

반면 사업자는 학습데이터 공개 의무화에 반대한다. 첫째, 데이터 공개는 경쟁기업에 영업비밀을 노출하는 결과로 이어질 수 있고, 데이터 경쟁 시대에 우위를 선점할 수 없어 경쟁력 상실로 이어질지 모른다는 우려 때문이다.¹⁷⁾

둘째, 현실적으로 생성형 AI에 사용되는 학습데

이터의 양이 방대해 기술적으로 데이터 공개가 어려울 수 있다. 이는 특히 대규모 언어모델의 경우, 수많은 출처에서 데이터를 수집하기 때문에 정확한 출처 추적이 어렵다.¹⁸⁾

셋째, 학습데이터 공개 의무화와 같은 규제가 국가의 산업 발전을 저해할 것이라는 우려가 있다. 특히 특정 국가들만 학습데이터 공개를 의무화한다면 해당 국가의 사업자는 글로벌 경쟁에서 불리한 위치에 놓이고, 투자 유인이 낮아지게 된다.¹⁹⁾

4 해외 입법 동향

(1) 미국

미국에서도 학습데이터 공개 의무화와 관련된 여러 입법 시도가 이어지고 있으며, 저작권 관련 법제보다 AI와 관련된 별도 법제를 통해 규율하려는 경향이 나타난다. 2024년 4월 발의된 「생성형 인공지능 저작권 공개법안(Generative AI Copyright Disclosure Act)」은 생성형 AI 학습에 이용된 저작물에 관한 통지 의무를 신설하여, 훈련 데이터를 생성하거나 변경하는 자가 출시 30일 전까지 '저작물 이용에 대한 구체적인 내용'을 저작권등록부에 제출하도록 규정하고 있다. 또한 저작권청은 제출된 정보를 바탕으로 공개 데이터베이스를 구축하고 이를 관리해야 한다.

이어 2024년 11월 발의된 「인공지능 네트워크의 투명성 및 책임에 관한 법안(Transparency and Responsibility for Artificial Intelligence Networks Act)」은 저작권자가 AI 개발자 또는 배포자를 상대로 저작물 학습 여부를 확인하기 위해 '확실하게 식별할 수 있는 사본 또는 기록'을 제출하도록 연방법원에 소환을 청구할 수 있는 절차를

13) 전우정·노태엽, 「인공지능 산출물과 학습데이터에 관한 저작권 문제 고찰」, 『저스티스』 제203호, 2024, p.209.

14) 김창화, 「생성형 AI를 둘러싼 최근 저작권 쟁점 및 판례 동향」, 『규제법 제리뷰』 제24-1호, 한국법제연구원, 2024, p.198.

15) 이철남, 앞의 글, p.37.

16) 대표적으로 Open AI는 'Financial Times', 'Shutterstock' 등과 저작물 사용에 대한 계약을 체결하였다고 밝혔다.

17) 구문모, 「생성형 인공지능의 데이터 이용과 저작권 쟁점에 관한 소고 - 음악저작물의 공정한 이용과 보상을 중심으로-」, 『제2회 음악저작권 관련 논문 공모전 수상집』, 한국음악저작권협회, 2024, p.12.

18) Yang, L. et al. "Difficulty aware programming knowledge tracing via large language models". *Scientific Reports*, No.11475, 2025, p.17.

19) OECD, "Artificial Intelligence in Society", Paris: OECD Publishing, 2019, p.100.

도입하는 것을 주요 내용으로 한다.²⁰⁾

다만 이러한 연방 차원의 입법 시도는 아직 본격적인 입법으로 이어지지 않았으며, 실질적인 규율은 주 단위에서 먼저 나타나고 있다. 대표적으로 캘리포니아주는 2024년 9월 「생성형 인공지능 훈련 데이터 투명성법(Generative Artificial Intelligence: Training Data Transparency)」을 제정하고, 2026년 1월 1일부터 시행할 예정이다. 이 법은 생성형 AI 개발자에게 모델 훈련에 사용된 데이터의 출처, 소유자, 데이터가 지식재산권으로 보호되는지 여부, 구매·이용허락 여부 등을 웹사이트에 공개하도록 의무를 부과하고 있다.²¹⁾

(2) 유럽연합(EU)

유럽연합(EU) 역시 학습데이터 공개 의무를 인공지능 관련 별도 법제를 통해 규정하고 있다. 유럽연합은 AI 규제를 위해 세계 최초로 「인공지능법(Artificial Intelligence Act)」을 제정하였으며, 이 법은 챗봇 등 범용 AI 모델의 훈련에 사용된 데이터 출처 공개를 의무화하고 있다. 2025년 8월부터 시행되는 관련 규정에 따라, 범용 AI 모델 제공자는 학습, 테스트, 검증에 이용된 데이터의 유형, 출처, 큐레이션 방법 등을 AI 사무국 및 국가 유관 기관에 제출해야 한다.²²⁾ 위반 시 일정 금액 또는 직전 회계연도 전 세계 매출을 기준으로 하는 벌금이 부과된다.²³⁾ 현재 회원국들은 법 규정에 명시된 데이터의 유형, 출처 및 제공 방식 등 구체적인 이행방안을 논의하고 있다.²⁴⁾

20) 손휘용, 「[미국] 인공지능 학습 과정에서 투명성 확보를 위한 미국의 정책 동향」, 『저작권동향』 제23호, 한국저작권위원회, 2024.

21) 김도경, 「인공지능에 의한 저작물 사용 관련 저작권 침해 책임 면책과 저작권자에 대한 보상에 관한 연구」, 『법조』 제73권 제5호, 2024, p.249. 및 김지현, 「미국의 인공지능 관련 주(州) 입법례」, 『최신외국입법정보』 제268호, 국회도서관, 2025, p.5.

22) 유럽연합 「인공지능법」 제53조(범용 AI 모델 제공자에 대한 의무)

23) 유럽연합 「인공지능법」 제99조(벌칙)

24) 정윤경, 「유럽연합 인공지능법(EU AI Act) 제정의 저작권법적 시사점」, 『IP & Data』 제4권 제2호, 2024, pp.86~87.

5 향후 입법 과제

학습데이터 공개 의무화가 사업자에게 과도한 규제가 될 수 있다는 우려가 있지만, 장기적으로 인공지능의 신뢰성 확보와 저작권 보호를 위해 논의 지속해야 한다. 다만 이해관계자의 의견이 대립하는 점을 고려하여 학습데이터 공개 의무 규정에 대한 단계별 입법이 이루어져야 하며, 「인공지능 기본법」과 「저작권법」에서 해결해야 할 과제를 구분하여 검토할 필요가 있다.

단기 과제로는 「인공지능 기본법」에 사업자가 권리자의 요청에 따라 저작물 학습 여부를 개별적으로 공개하는 제한적 공개 규정을 마련할 수 있다. 「인공지능 기본법」 제4장 ‘인공지능 윤리 및 신뢰성 확보’에 ‘인공지능 투명성 확보 의무(제31조)’가 규정되어 있는 만큼, 이를 보완하여 제한적 공개 의무를 「인공지능 기본법」 제31조 내 사업자의 책무로 신설하는 방식이 「저작권법」에 인공지능 사업자에 관한 별도의 의무 규정을 신설하는 것보다 법체계의 정합성 측면에서 더 적절한 것으로 보인다.

장기 과제로 학습데이터 전체 공개를 의무화하는 규정을 「인공지능 기본법」에 신설하는 방안도 고려할 수 있다. 다만, 이는 제한적 공개 규정 도입 이후의 효과와 AI 산업 발전 동향을 종합적으로 검토하고, 충분한 사회적 합의가 이루어진 상황에서 논의하는 것이 바람직하다.

한편, 「저작권법」에서는 관련 입법 동향과 국내 산업 실무를 반영하여 저작물 이용에 관한 저작권 면책 규정의 도입방안을 검토해야 한다. 아울러, 이에 대한 보상체계 마련 역시 이해관계자의 의견을 수렴하여 사업자와 권리자 간 이익 균형을 달성할 수 있는 방향으로 추진할 필요가 있다.

『이슈와 논점』은 국회의원의 입법활동을 지원하기 위해 최신 국내외 동향 및 현안에 대해 수시로 발간하는 보고서입니다.

