

생성형 AI 시대 저작권 보호를 위한 학습 불가능한 예제와 기술적 보호조치의 개선 방안 연구

김 중 성^{1†} · 박 성 환^{2‡} · 황 요 한³ · 오 병 훈³ · 이 재 우^{4*} · 홍 준 호^{5*}

¹중앙대학교 융합보안학과 연구교수

²중앙대학교 융합보안학과 박사수료

³중앙대학교 융합보안학과 석사과정

⁴중앙대학교 산업보안학과 부교수

⁵성신여자대학교 융합보안공학과 조교수

A Study on the Improvement of Unlearnable Examples and Technical Protection Measures for Copyright Protection in the Generative AI Era

Jongseong Kim^{1†} · Sunghwan Park^{2‡} · Yohan Hwang³ · Byunghoon Oh³ · Jaewoo Lee^{4*} · Junho Hong^{5*}

¹Research Professor, Department of Convergence Security, Chung-Ang University, Seoul 06974, Korea

²Master's Course, Department of Convergence Security, Chung-Ang University, Seoul 06974, Korea

³Master's Course, Department of Convergence Security, Chung-Ang University, Seoul 06974, Korea

⁴Associate Professor, Department of Industrial Security, Chung-Ang University, Seoul 06974, Korea

⁵Assistant Professor, Department of Convergence Security Engineering, Sung-shin University, Seoul 02844, Korea

[요 약]

본 연구는 생성형 AI(Artificial Intelligence)의 학습 데이터 활용 과정에서 발생하는 저작권 침해 쟁점을 분석하고 공정이용 법리의 불확실성과 아이디어·표현 이분법의 한계를 보완하기 위한 기술적·제도적 개선방안을 제시한다. 특히 기술적 보호조치로 학습 불가능한 예제 기법의 실효성을 검증했다. 결론적으로 「저작권법」상 기술적 보호조치로서의 기능을 재정립하고 권리자 옵트아웃 시스템과 연계하는 이중 주체 기반 보호 프레임워크를 제안했다. 이를 통해 AI 학습의 법적 불확실성을 해소하고 저작권자의 창작 유인 도모 및 AI 산업의 지속 가능한 발전을 위한 책임 기반을 마련하는 데 기여하고자 한다.

[Abstract]

This study analyzes the institutional and legal gaps that arise in the Copyright Act when generative AI systems learn from copyrighted works and other data without explicit authorization. Comprehensive measures are proposed to address these deficiencies. By conducting a comparative analysis on the copyright laws of major jurisdictions, including text and data mining (TDM) exceptions and recent case law, the study identifies critical limitations within the current domestic legal framework in regulating AI training practices. Furthermore, through empirical validation, the research examines the feasibility of the unlearnable examples technique as a practical protection measure to safeguard copyrighted materials during the AI learning process. Building on these insights, the study ultimately proposes complementary institutional and policy reforms designed to establish a balanced ecosystem that ensures both effective copyright protection and sustainable advancement of generative AI innovation.

색인어 : 저작권 보호, 학습데이터, 기술적 보호조치, AI, 학습 불가능한 예제

Keyword : Copyright Protection, Training Data, Technical Protection Measures, Artificial Intelligence, Unlearnable Examples

<http://dx.doi.org/10.9728/dcs.2025.26.12.3507>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 30 October 2025; **Revised** 10 November 2025

Accepted 14 November 2025

‡ These authors contributed equally to this work

*Corresponding Author; Jaewoo Lee and Junho Hong

Tel: +82-2-8742-3738

E-mail: hjh@sungshin.ac.kr

I. 서 론

생성형 AI(Artificial Intelligence, 이하 AI)의 급속한 확산과 함께 저작물 등을 포함한 대규모 학습데이터를 AI 개발에 활용하는 행위의 적법성에 관한 논의가 심화되고 있으며 특히 이른바 ‘지브리풍’ 이미지 논쟁 등 사례가 대중적으로 확산되면서 쟁점의 긴급성이 커지고 있다. 우리 「저작권법」은 제35조의5(공정이용)를 통해 이용 목적·성격, 저작물의 종류·용도, 이용의 양과 중요성, 시장에 미치는 영향 등 4요소를 종합 고려하도록 하고 있어 AI(Artificial Intelligence, 이하 AI) 서비스 제공자는 이를 학습 단계 정당화 논리로 사용하나 권리자 측은 무단 복제·공중송신에 의한 명백한 권리 침해로 본다는 대립 구도가 형성되어 있다. 더구나 우리나라는 미국과 달리 AI 학습 단계의 공정이용 적용을 직접적으로 판단한 판결이 아직 선례로 축적되어 있지 않아 법적 예측가능성이 낮은 반면, 미국에서는 최근 판결들이 공정이용의 제1요소부터 제4요소를 종합적으로 판단하는 경향을 분명히 하고 있어 향후 우리 법원의 중요한 참조 틀로 기능할 가능성이 높다. 비교법적으로는 일본이 2018년 저작권법 개정을 통해 비향수목적의 텍스트·데이터 마이닝(Text and Data Mining, 이하 TDM)에 광범위한 면책을 부여하고 경미한 이용에 대한 별도 면책을 둬으로써 학습데이터 활용에 비교적 유연한 체계를 운용하는 한편, EU(European Union, 이하 EU)도 디지털 단일 시장 지침(Digital Single Market, 이하 DSM 지침)을 통해 연구기관 특허(제3조)와 일반 TDM(옵트아웃 전제, 제4조) 체계를 이원적으로 설계하고 2024년 AI Act에서 범용 AI 모델 제공자에게 권리자 옵트아웃 확인 및 학습내용 요약 의무를 부과하는 등 제도화를 진전시켰다.

반면 TDM 면책 조항이 없는 우리나라는 당분간 미국 판례 전개를 참고한 공정이용 해석에 의존할 수밖에 없을 것으로 보이며 ‘스타일’·‘화풍’과 같은 아이디어 영역을 원칙적으로 보호하지 않는 「저작권법」의 전통적 한계인 아이디어·표현 이분법 원칙과 학습단계 적법성의 불확실성이 중첩되는 구조적 공백이 지속되고 있다[1]. 본 논문은 이러한 공백을 메우기 위해 기술 및 제도적 측면에서 권리자가 시각적 품질을 유지하면서도 AI의 무단 학습을 실질적으로 방해하도록 데이터 자체의 학습 유용성을 파괴하는 ‘학습 불가능한 예제’(unlearnable examples)에 관한 실험을 실시했다. 또한 온라인 서비스 제공자 의무 규정의 체계를 참고하여 AI 서비스 제공자에게 저작권자가 본인의 저작물의 유통을 방지할 수 있도록 의사를 표시하면 기술적 보호조치를 시행해야 하는 의무를 명확히 부과하는 법제 프레임워크를 제안한다. 이를 통해 저작권자와 AI 서비스 제공자 간 공정이용의 불확실성을 보완하는 것이 본 연구의 목적이다.

II. 생성형 AI와 학습데이터 활용 및 보호

2-1 생성형 AI와 저작권 충돌: 데이터 기반 AI 학습과 법적 공백

생성형 AI는 대량의 데이터를 학습하여 AI가 만든 결과물을

도출하고 있다. 이는 생성형 AI가 데이터 학습에 의존하고 있다는 것을 확인할 수 있다. 그러나 AI시스템이 데이터로 학습하는 단계에서 해당 데이터에 대해 저작권이 있을 경우 저작권 침해에 대한 문제가 제기될 수 있다.

문화체육관광부와 한국저작권위원회가 2023년 12월에 발간한 생성형 AI 저작권 안내서는 생성형 AI를 대규모 데이터셋에 기반한 딥러닝 기술을 활용하여 이용자가 요구하는 새로운 데이터 또는 콘텐츠 등의 산출물을 제시하는 AI 기술이라고 정의한다[1]. 즉 생성형 AI는 다양한 딥러닝 기반의 생성 모델을 통해 이용자의 요구에 따라 텍스트, 이미지, 오디오 등의 생성 작업을 수행하는 AI 전반을 포괄하는 개념으로 설정할 수 있다[2]. 딥러닝은 합성곱 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN) 등 신경망 구조를 통해 다양한 유형의 데이터를 처리 및 분석할 수 있어 이를 기반으로 학습하는 생성형 AI의 성능은 데이터의 양과 질에 비례하여 결정된다[3]. 생성형 AI인 GPT(Generative Pre-trained Transformer, 이하 GPT)-3.5의 경우 데이터 4,100억 개, 논문·보고서 등 웹 텍스트 190억 개, 인터넷 기반의 말뭉치 670억 개 등을 학습했으며 이후 출시된 ChatGPT-4는 GPT-3보다 매개변수를 기준으로 10배 이상 큰 만큼 학습량이 방대하다[4]. 생성형 AI의 데이터 학습은 기존의 저작물 등을 포함한 데이터의 대량 복제가 수반되는 것이라 할 수 있으며 AI 사업자 입장에서는 수많은 데이터를 사용하므로 각 권리자에게 이용허락을 개별적으로 구하는 것은 어렵다[5].

이러한 현실적인 문제로 인해 생성형 AI의 데이터 학습 과정은 데이터셋을 구성하는 데이터가 저작물일 경우 「저작권법」상 권리자의 복제권 및 공중송신권 등을 침해할 수 있어 권리자와 AI 사업자 간 분쟁의 여지가 다분하다. AI시스템이 학습하는 과정에서 저작권이 있는 데이터를 학습하는 경우 복제권 침해에 대한 문제가 제기될 수 있다. 여기서 복제권 침해 여부와 관련하여 「저작권법」 제35조의2(저작물의 이용과정에서의 일시적 복제)의 논리를 기반으로 효율적인 정보처리를 위해 데이터셋을 일시적으로 복제했다고 할 수 있다. 그러나 「저작권법」 제35조의2를 포함하여 동법 제23조부터 제35조의4까지의 규정은 개별적인 지식재산권 제한사유로 비영리적 또는 공익적인 목적으로 타인의 저작권을 이용할 수 있도록 허용한다[6]. 생성형 AI는 데이터 학습을 통해 새로운 가치를 창출하여 영리적인 목적을 달성하고자 하므로 위 규정들에서 열거하고 있는 개별적인 사유에 해당하지 않아 일시적인 데이터셋의 복제는 지식재산권 제한사유에 해당하지 않는다. 생성형 AI의 데이터 기반 학습은 권리자가 보유한 이익을 훼손하지 않는 범위 내에서 이루어져야 할 것인 바, 「저작권법」 제35조의5에 명시된 공정이용 규정이 AI 사업자와 권리자 간 이해관계를 조율할 근거로 활용될 수 있다. 공정이용이란 저작권자의 허락 없이 저작물을 이용할 수 있는 저작권에 대한 특수한 경우를 말한다.

「저작권법」 제35조의5는 미국 「저작권법」 제107조 공정이

용 규정을 토대로 한 조항으로 ① 이용의 목적 및 성격 ② 저작물의 종류 및 용도 ③ 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성 ④ 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향을 고려하여 공정이용 여부를 판단한다. 본 조항은 네 가지 요소를 통해 공정이용의 범위를 포괄적으로 규정하고 있어 AI 등 신기술의 발달에 따라 공정이용의 적용 가능성을 폭넓게 논의할 수 있다는 장점이 있다. 우리나라 현행 「저작권법」상 공정한 저작물 등을 활용하기 위해 공정이용 제도를 도입하여 시행하고 있다. 현행 「저작권법」상 공정이용에 대해 판단할 시 전통적 이용인 교육, 연구, 패러디 등에 있어 공정이용에 대한 판단요소들을 정립되어 있으나, AI시대 AI시스템이 데이터를 학습하는 과정에서 공정이용 관점을 적용하기는 어려운 상황이다.

따라서 AI 시스템이 데이터를 학습하는 과정에서의 쟁점들을 분석할 필요가 있다. 먼저 AI시스템이 데이터를 학습하는 과정에 있어 ‘이용 목적 성격’에 대해 살펴보면, AI시스템이 데이터를 학습하는 것은 일반적으로 구글, 오픈AI처럼 상업적 기업이기 때문에 상업적 활용을 위해 데이터를 학습하는 행위는 공정이용의 판단요소인 ‘이용 목적·성격’상 상업적 행위이기 때문에 공정이용 범위에 포함되지 않는다. 그러나 기술 발전과 공익적 연구 기여 등을 위해 활용한 AI시스템의 경우 공정이용에 있어 비영리 및 연구 목적이기 때문에 공정이용 범위에 포함될 수 있다. 두 번째로 ‘저작물의 종류와 용도’이다. 생성형 AI 시스템은 학습된 데이터 및 입력 데이터에 따라 도출되는 결과가 상이하다. 그로 인해 저작물의 종류와 용도를 판단하는 명확한 기준을 세우기 어렵다. 그로 인해 저작물의 종류와 용도는 생성형 AI시스템이 학습된 데이터를 기반으로 도출된 결과물에 따라 저작물의 종류와 용도는 서로 달라지게 된다. 세 번째로 이용의 양과 중요성이다. 종래 「저작권법」상 저작물을 일부 인용은 공정이용으로 보고 있으나, 저작물의 전체를 사용하게 되면 공정이용 인정 여부가 달라지게 된다. 그러나 생성형 AI시스템에 있어 AI가 학습하는 데이터는 작품의 전체가 학습되며, 전체 학습된 데이터 중 일부분 등을 활용하여 결과물을 도출할 수 있기 때문에 이용의 양과 중요성에 대해 판단 시 명확한 기준과 판단 요소가 없는 것이 현실이다. 마지막으로 시장에 미치는 영향이다. 기존 「저작권법」상 저작권자의 시장을 대체하는 경우, 이는 공정이용으로 볼 수 없다. 그러나 생성형AI 시스템이 생성된 결과물이 원저작물과 유사한 저작물인지 또는 대체재로 활용될 수 있는지 여부 등에 따라 시장에 미치는 영향이라는 판단요소를 적용할 수 있을 것이다.

현재 우리나라에서는 공정이용 조항을 다룬 판결은 2024년 11월에 선고된 2건의 판결만 존재하고 있다. 또한 AI 학습에 있어 공정이용 규정의 적용 여부를 판단한 국내 법원의 판례 또한 전무하여 우리나라의 「저작권법」상 공정이용의 범위가 사실상 명확하지 않은 실정이다.

반면 최근 미국에서는 AI의 개발을 위해 저작물을 학습시키는 행위가 공정이용 여부에 해당하는지에 관한 판결들이 등장하고 있다. 미국과 달리 우리나라는 공정이용에 대한 다수의

판례 등이 부재하여 공정이용에 대한 명확한 법리 해석을 위한 기초적 자료가 부재한 한계가 존재하고 있다. 따라서 공정이용을 판단하기 위한 4가지 요소들은 AI 시스템이 데이터를 학습하는 과정에서 해석상의 어려움과 제도적인 판단의 공백들이 존재하고 있다. 이에 현행 「저작권법」상 공정이용 적용 여부는 아직까지도 불확실하며, 법적 공백으로 인해 AI산업 발전을 물론 학습데이터를 제공하는 저작권자 간의 불균형들이 아직까지도 존재하고 있다. 이는 법적 공백으로 인해 데이터 학습을 전제로 하는 생성형 AI와 저작권 보호 체계 간의 충돌을 야기하고 있음을 시사하고 있다.

2-2 「저작권법」의 전통적 한계: 아이디어·표현 이분법 원칙

특허권과 저작권은 정신활동의 산물인 지적 창작물을 보호한다는 공통점을 가지고 있다. 그러나 권리객체를 중심으로 살펴보면 전자는 기술적 사상과 같은 아이디어 자체를, 후자는 아이디어의 ‘표현’을 보호한다는 점에서 구별된다. 「저작권법」 제2조는 저작물을 인간의 사상 또는 감정을 표현한 창작물로 정의하므로 보호 대상은 사상 또는 감정에 대한 아이디어가 아니라 창작적 표현으로 이른바 아이디어·표현 이분법이 저작권 체계를 관통하는 기본 원칙이다. 아이디어는 공중의 영역에 속하므로 독점·배타적 권리를 부여할 수 없어 제3자가 저작물에 내재된 아이디어를 무단으로 사용하더라도 구체적 표현을 모방하지 않았다면 저작권 침해가 성립하지 않는다[7]. 아이디어와 표현의 개념에 한 요소로 작용하는 사상과 감정은 철학적이거나 심리학적인 개념으로 좁게 해석할 것이 아니라 생각이나 기분의 정도를 포함하는 넓은 의미로 해석해야 하며 자연과학적 사실이나 사회적 사실 등 단순한 사실은 사상이나 감정의 표현이 될 수 없다[8]. 아이디어·표현 이분법은 장기간의 판례 축적을 통해 발전하여 미국을 비롯한 다수 국가에서 저작물 보호범위를 정하는 기본 법리로 기능하고 미국 「저작권법」뿐 아니라 WTO (World Trade Organization, 이하 WTO)/TRIPS (Agreement on Trade-Related Aspects of Intellectual Property Rights, 이하 TRIPS), WIPO(World Intellectual Property Organization, 이하 WIPO) 저작권조약 등에서도 표현 중심의 보호 원칙을 명문으로 규정하고 있다[8].

다만, 「저작권법」의 보호범위에서 제외되는 아이디어의 정의를 명확히 내리는 것은 어려우며 아이디어와 표현의 경계는 사안별로 달라질 수 있다[9]. 미국 법원의 경우 수많은 판례의 집적으로서 아이디어와 표현의 이분법을 발달시켰으며 특히 미국 「저작권법」 제102조(b) 규정에 따라 아이디어를 세 가지로 분류한 Goldstein 교수의 개념 정의가 널리 받아들여지고 있다. Goldstein은 아이디어와 표현의 구분 작업에 있어 현재와 미래의 창작자 이익을 균형 있게 고려할 것과 구분이 곤란한 경우 아이디어를 밝히기보다 보호되는 표현의 범위를 정하는 것을 중요한 원리로 제시한다. 결국 아이디어와 표현의 구별은 정도의 문제로 볼 수 있으며 어느 지점부터 둘의 개념이 구분되어 보호가 되는 것인지 경계를 설정하는 것은 상당히 어

렵다[10].

한편, 아이디어와 표현의 경계가 극히 밀접하게 합체되어 서로 분리할 수 없게 될 정도로 모호한 선상에 놓이는 경우가 있다. 이러한 표현을 보호하는 것은 결과적으로 그 아이디어까지 보호하게 되는 것이므로 아이디어·표현 이분법 원칙을 반하여 저작권의 보호를 주지 않는 것을 합체의 원칙이라 한다. 또한 저작물을 창작한 시점에 아이디어를 표현할 방법이 있었으나 기술의 발전 등 요인으로 아이디어와 표현의 경계가 모호해질 경우에도 합체의 원칙에 따라 「저작권법」상 보호를 받을 수 없다.

종합하면 「저작권법」은 아이디어·표현의 이분법 및 합체의 원칙 등을 통해 사상 또는 감정 그 자체에 해당하는 아이디어가 아니라 이를 표현한 창작물에 한하여 보호한다는 점을 명확히 하고 있다. 그러나 생성형 AI 기술 발전으로 인해 AI가 만들어 내는 산출물들은 특정 작가의 화풍, 서사 구조, 작곡 스타일 등을 학습하는 등 이러한 행위들은 기존의 ‘아이디어’ 개념과 경계가 모호해지는 지점들이 나타나고 있다.

2-3 저작권 침해와 아이디어의 표현 쟁점 사례 분석

1) 게임의 기본규칙과 게임에서 구현된 표현에 관한 판결

2014년 2월 국내 게임회사인 아보카도 엔터테인먼트(이하 ‘아보카도’)는 영국 게임회사인 킹닷컴 리미티드(이하 ‘킹닷컴’)의 게임이 큰 인기를 끌자 이와 유사한 게임을 출시하였다. 킹닷컴은 양 게임의 전개 방식, 규칙, 인터페이스, 캐릭터, 게임 보드의 구성과 배치 등 게임의 시각적 디자인이 유사하다는 이유로 「저작권법」상 저작권 침해 및 「부정경쟁방지법」 차목의 부정경쟁행위에 해당한다고 주장하며 아보카도를 상대로 손해배상을 청구했다[11]. 본 장은 인간의 사상 또는 감정에 해당하는 아이디어와 창작적 표현의 보호범위를 「저작권법」의 관점에서 살펴보기 위해 마련한 것이므로 「부정경쟁방지법」상 판결의 요지는 제외하였다. 해당 판결은 「저작권법」상 보호 대상이 표현의 선택에 관한 아이디어 자체가 아닌 구체적인 표현이므로 게임 화면의 구성 및 규칙, 게임 보드 구성 등에 선택의 여지가 있어도 실제 작성된 표현이 흔하게 사용되는 것이라면 저작권 침해에 해당한다고 볼 수 없다고 하였다. 즉 표현의 선택에 관한 아이디어는 동일해도 구체적인 표현 자체에 차이가 있다면 저작권 침해에 해당한다고 볼 수 없다고 하여 아이디어·표현 이분법이 적용돼 「저작권법」의 보호를 받을 수 없다고 판결된 사건이다.

2) 오픈AI의 지브리풍 이미지 학습, 저작권 침해 논란

2025년 3월 오픈AI는 챗GPT-4o 이미지 생성 모델을 출시하면서 기존 이미지 모델인 ‘달리(DALL-E)’보다 쉽고 정교하게 이미지를 생성한다고 소개했다. 과거에는 이용자가 원하는 이미지를 위해 프롬프트를 일일이 입력해야 하지만 새 모델은 이용자의 의도를 파악하여 간편하게 다수의 이미지를 생성해 낼 수 있게 되었다. 챗GPT-4o는 지브리와 디즈니 등 다양한 애니메이션 화풍의 이미지를 생성해 낼 수 있어 대중들은 일상

사진을 업로드하고 지브리풍으로 그려달라는 간단한 프롬프트를 입력해 완성도 높은 애니메이션 스타일의 이미지를 얻을 수 있게 되면서 무려 5억 명의 신규이용자가 증가하였다[12]. 한국은 물론 전 세계적으로 지브리 스타일 이미지에 대한 관심이 폭증하였으나 지브리 스튜디오의 아버지인 미야자키 하야오 감독은 AI가 만든 애니메이션 또는 이미지에 부정적인 의견을 드러냈다[13]. AI기술의 발전과 생성형 AI 등장 이후 AI 창작물에 관한 논의는 크게 두 가지로 나뉜다. 먼저, AI 창작물의 저작물성이 인정될 수 있는지 여부는 인간의 창작이 아니므로 그 저작물성을 부정하는 것은 정립되었다고 보는 시각이 존재한다[14]. 또한, AI에 입력하는 프롬프트 쟁점 중 하나로 새벽의 자리아 사건처럼 AI 창작물을 인간이 편집한 경우의 저작물성 등에 대한 쟁점들이 꾸준히 제기되고 있다. 새벽의 자리아의 경우 미국 저작권청(United States Copyright Office, 이하 USCO)이 2023년 저작권 등록을 한 사건으로 생성형 AI가 만든 그림을 사람이 선택, 배열한 것에 편집저작물로 등록 가능성을 보여줬다[15]. 챗GPT-4o가 지브리풍 등 스타일을 학습하여 결과물을 산출한 것은 사람이 만든 창작물이 아니면서 별도의 선택, 배열 등 가미된 요소가 없어 저작물성은 없다고 보는 것이 타당해졌다. 한편, AI가 타인이 창작한 저작물 데이터를 학습할 경우 저작권 침해인지에 대한 여부는 「저작권법」상 공정이용의 법리를 활용해 판단해 볼 수 있었으나 지브리풍 사건의 경우 아이디어·표현 이분법 원칙에 따라 저작권 침해가 성립되지 않는다. 스타일이라 함은 회화에서 특정한 스승과 제자 그룹 간에 전승되는 유(流) 혹은 인상주의, 표현주의, 입체주의와 같은 유파(流派)나 랩트란트풍(風), 모네풍(風) 등으로 인간의 사상 또는 감정의 아이디어의 개념에 포함되기 때문이다[16]. 지브리풍 사건으로 AI 학습 저작권 논란이 확산되었고 관련 스타일을 무단으로 활용해 상업적 목적으로 이익을 취하여 권리자의 실질적인 금전적 피해 등을 살피는 「부정경쟁방지법」상 처벌은 고려해볼 수 있을 것이나, 아이디어에 해당하는 스타일을 생성형 AI가 학습하여 결과물을 낸 것은 기본적으로 저작권 침해가 성립되지 않는다.

2-4 시사점

생성형 AI는 딥러닝 기반으로 운용되는 특성상 학습 데이터의 규모와 다양성이 성능을 좌우하므로, 사업자는 GPT-3.5의 사례와 유사하게 크롤링 등으로 대량의 데이터를 수집·전처리하여 학습시키게 된다. 이 과정은 기존 저작물 등을 포함한 데이터의 대량 복제를 수반할 수 있어 저작권자의 복제권·공중송신권 침해 가능성이 제기되며 방대한 권리관계의 특성상 개별 권리자로부터 일일이 이용허락을 받는 것은 현실적으로 곤란하다.

따라서 「저작권법」 제35조의5의 공정이용 규정이 이해관계 조정의 근거로 기능할 여지가 있으나, 현재까지 국내에서 공정이용을 근거로 AI 학습 단계의 적법성을 직접 판단한 판결은 전무하여 적용 범위가 불확실하고, 이 같은 법적 공백은 데

이터 학습을 전제로 하는 생성형 AI와 저작권 보호 체계 간의 구조적 충돌로 이어지고 있다. 더불어 저작권법은 저작물을 ‘인간의 사상 또는 감정을 표현한 창작물’로 한정하므로 아이디어 자체는 보호하지 않는 아이디어·표현 이분법을 전제하고 실제로 게임의 기본규칙 및 전개 방식과 같이 ‘표현의 선택에 관한 아이디어’가 동일하더라도 구체적 표현이 상이하면 침해가 성립하지 않는다는 취지가 판례를 통해 확인됐다. 오픈AI의 이른바 지브리풍 사례 역시 기본적으로 특정 ‘스타일’이 아이디어 영역에 속한다는 점에서 저작권 침해로 평가되기 어렵고, 다만 상업적 이용과 결합한 부정경쟁방지법상 쟁점은 별도로 논의될 수 있다. 물론 공공의 영역에 속하는 아이디어나 스타일을 곧바로 저작권 보호범위에 편입하자는 취지는 아니나 누가 보아도 특정 출처를 즉시 연상시킬 정도로 구체적 특징이 축적되는 경계 사례에 대하여는 ‘합체의 원칙’상 표현을 보호하는 것이 사실상 아이디어 보호로 귀착될 위험을 면밀히 점검할 필요가 있다.

요컨대 현행 체계하에서는 공정이용의 해석 불확실성과 아이디어·표현 이분법의 경계 문제가 중첩되어 학습 단계의 법적 안정성이 확보되지 못하는 한계점이 존재한다. 이는 저작권자의 창작 유인을 약화시켜 「저작권법」 제1조의 목적에도 반하는 결과를 초래할 우려가 있으므로, AI시대 저작권법의 공백을 해결할 수 있는 방안의 관해 모색할 시점이다.

III. 주요국 AI와 학습데이터 보호조치 현황 분석

3-1 미국

미국은 저작물 등을 포함한 학습데이터의 보호와 활용 사이의 균형을 맞추기 위해 연방 「저작권법」(17 U.S.C. § 107)을 근거로 공정이용 여부를 판단한다. 연방 「저작권법」 제107조는 논평, 뉴스, 교육, 연구 등의 목적을 위하여 저작권자의 동의나 허락 없이 원저작물을 이용하는 행위가 공정이용에 해당하는지를 네 가지 고려 요소로 정하고 있다. 제1요소인 ‘원저작물 이용의 목적과 성격’은 크게 상업적 이용 여부와 변형적 이용 여부로 나뉜다. 상업적 이용 여부를 고려하는 것은 원저작물의 이용에 대해 대가를 지급하지 않으면서 그로부터 상당한 이익을 얻는 것이 형평성에 어긋난다는 점에 근거한다[17]. 변형적 이용은 원저작물을 단순히 대체하는 수준을 넘어 원저작물에서 찾아볼 수 없는 새로운 표현이나 의미 등을 추가하여 원저작물과 다른 별개의 목적이나 성격을 갖게 되는 경우를 의미한다[18]. 이때 공정이용 판단에서 중요한 것은 원저작물의 내용이나 형태를 실제로 바꾸었는지보다 원저작물을 기존과 다른 목적으로 이용했는지에 대한 판단 여부를 살펴보고 있다.

Thomson Reuters v. Ross Intelligence 사건은 제1요소가 쟁점이 된 사례로 피고 로스(Ross)는 기존 법률 문서를 검색 및 제공하는 AI를 개발하기 위해 ‘웨스트로(Westlaw)’의 법률 판례 요약물 무단 복제해 학습에 활용했다[19]. 법원은 로스의 서비스가 새로운 창작물을 생성하지 않고 원고의 법률

검색 기능을 모방하는 데 그친다고 보아 해당 이용을 비변형적 이용으로 판단하고 공정이용을 부정하였다. 한편, 제2요소인 ‘저작물의 종류 및 용도’와 관련해서는 저작물이 사실에 기반하기 보다 창작성이 높은 표현 위주일수록 공정이용으로 인정되기 어렵다는 것을 나타낸다. 제3요소인 ‘이용된 부분이 저작물 전체에서 차지하는 비중과 중요성’은 저작물의 이용이 목적 달성에 필요한 수준을 넘어서는 안 되며 이용된 부분의 양은 이용하고자 하는 목적과 관련하여 합리적인 수준이어야 한다. 어느 정도의 이용이 목적 달성에 필요하고 합리적인 수준인지 문제될 수 있으나 원저작물의 전체를 복제하는 경우에도 다른 요소들을 고려한 결과 공정이용으로 인정될 수 있다[20]. 제4요소인 ‘저작물 이용이 그 저작물의 시장 또는 가치에 미치는 영향’은 이용행위가 원저작물을 대체하거나 그 시장 가치에 심각한 타격을 줄 경우 공정이용 인정이 어려움을 뜻한다. 장래에 손해가 발생할 상당한 가능성이 인정되는 경우에도 제4요소는 공정이용 성립에 불리하게 작용한다.

Bartz v. Anthropic 사건은 제1요소부터 제4요소까지를 종합적으로 검토했을 뿐 아니라 학습에 사용된 저작물의 적법성도 함께 다룬 점에서 의미가 있다. 피고 앤트로픽(Anthropic)은 수백만 권의 서적을 구매하고 불법복제 서적도 활용하여 자사의 거대 언어모델(Large Language Model, 이하 LLM) ‘클로드(Claude)’를 학습시켰는데 법원은 구매 서적을 이용한 부분에 대해서는 공정이용에 해당한다고 보았다. 학습에 이용된 책들이 표현적 요소를 포함하고 있으나(제2요소) 이용은 완전히 변형적이었고(제1요소) 그 목적상 서적 전체를 이용하는 것이 합리적이며(제3요소) 저작물의 잠재적 시장이나 가치에 미치는 영향 측면에서도(제4요소) 공정이용이 인정된다고 판단이다[21]. 다만 불법복제 서적의 이용은 공정이용에 해당하지 않는다고 보았고 2025년 9월 앤트로픽은 AI 학습 관련 사건 중 최대 규모로 약 2조 원의 손해배상에 합의하였다[22].

Kadrey v. Meta 사건은 제4요소가 특히 부각된 사례인데 피고 메타(Meta)는 토렌트 방식으로 수집한 데이터셋을 활용해 거대 언어 모델(LLM) ‘라마(LLaMA)’를 학습시킨 행위가 문제 되었으나 법원은 이 사건에서 공정이용을 인정하였다[23]. 제1~3요소에 관한 판단은 앞선 앤트로픽 사건과 유사하다고 보면서도 향후 다른 사건에서는 다른 결론이 나올 수 있음을 지적하였다. 또한 제4요소가 공정이용 판단에서 핵심적 요소임을 강조하며 극히 적은 시간과 창의성으로 무수히 많은 경쟁작을 생성할 수 있는 저작물의 생산은 원저작물 시장에 미치는 영향에 따라 저작권자에게 대가 지급 없이 생성형 AI 학습에 저작물을 사용하는 행위가 위법해질 수 있음을 시사했다.

요컨대 미국은 AI 개발을 위한 학습데이터 활용과 관련해 TDM 면책 규정을 별도로 두고 있지 않으나 연방 「저작권법」의 공정이용 조항에서 제1~제4요소를 개별적 관점에서 판단하는 것이 아닌 종합적으로 고려함으로써 저작권자의 권익 보호와 학습데이터를 활용한 AI 개발 활성화 간의 균형을 도모하려는 경향을 보인다.

3-2 EU

유럽연합(EU)은 AI가 저작물 이용과 관련한 학습데이터를 활용하는 것에 대해 2019년 「디지털 단일 시장 저작권 지침」(Directive on Copyright in the Digital Single Market, 이하 DSM 지침)을 제정해 TDM 면책 규정을 마련하였다. 이는 디지털 환경에서 나타나는 새로운 유형의 저작물 이용에 대응하기 위한 것으로 TDM 행위의 면책 요건을 제3조와 제4조로 이원화해 규정한다[24]. 먼저 제3조는 ‘특정 TDM 면책’ 조항으로 연구기관과 문화유산기관이 학술연구 목적의 TDM을 수행하기 위해 적법하게 접근 가능한 저작물 등을 복제 및 추출하는 경우를 면책 범위로 정한다(제1항). 즉 면책의 적용 대상을 연구기관과 문화유산기관으로 하고 목적을 학술연구로 제한하는 것인데 계약이나 저작권자의 유보(Opt-out) 등 조항의 적용을 배제할 수 없다는 점에서 강행적 효력을 갖는 것이 특징이다. 제4조는 ‘일반 TDM 면책’ 조항으로, 제3조와 달리 학술연구라는 목적 범위나 연구기관 등 실행주체의 제한 없이 적법하게 접근 가능한 저작물 등에 대한 TDM을 위해 필요한 복제·추출을 허용한다(제1항). 복제물과 추출물 보유의 범위와 조건도 제3조에 비해 완화된 TDM 목적 달성에 필요한 범위로 규정한다(제2항).

다만 본 조에 따른 TDM 면책은 온라인으로 공중이용 가능한 콘텐츠에 관하여 권리자가 기계판독 가능한 수단 등 적절한 방법을 통해 명시적으로 유보하지 않았을 것을 조건으로 한다(제3항). 따라서 온라인 콘텐츠와 같은 학습데이터에 대해 권리자가 기계가독형 수단을 통해 명시적으로 유보 의사를 표시한 경우에는 TDM 면책이 적용되지 않는다.

이러한 입장은 2024년 제정된 「AI 법」(AI Act)에도 반영되어 범용 AI 모델 제공자는 DSM 지침 제4조 제3항에 따라 권리자가 명시한 권리 유보를 확인해야 한다[25]. 아울러 범용 AI 모델 제공자에게는 AI 학습에 사용된 기술적 문서, 콘텐츠 등 내용물에 관한 상세한 요약 제공 의무가 신설되어 권리자가 DSM 지침에 따른 유보 의사를 보다 효과적으로 표명할 수 있도록 규정이 보완되었다[26].

요컨대 EU는 DSM 지침을 통해 TDM 면책의 적용 주체와 목적을 층위별로 설계하고 「AI 법」을 통해 범용 AI 모델 제공자에게 권리자 옵트아웃 확인 의무 등을 부과함으로써 학습데이터의 합법적 활용을 보장하는 한편 창작자 권익 침해를 최소화하려는 균형을 모색하고 있다.

3-3 일본

일본은 2018년 「저작권법」 개정을 통해 AI 개발을 위한 기계학습에 수반되는 데이터 복제 등과 관련한 저작권 침해 면책 규정을 동법 제30조의4와 제47조의5에 마련하였다. 제30조의4는 자신 또는 타인이 저작물에 표현된 사상이나 감정의 향수를 목적으로 하지 않는 경우 적용되는 면책 규정이다. 향수(享受)란 저작물을 인식하고 지적 또는 정신적 욕구를 충족시키는 이용을 가리키는 것을 뜻한다. 저작물의 경제적 가치는 저작물

에 표현된 사상과 감정을 향수하는 것의 대가이며 그 향수를 목적으로 하지 않는 이용은 형식적으로 저작물의 이용일지라도 「저작권법」이 보호하는 이용 형태로 볼 수 없다[27]. AI가 학습데이터의 활용을 위해 저작물을 이용하더라도 작품이나 화풍 등 아이디어의 유사성만 있을 뿐 기존 저작물과 표현상 유사성이 인정되지 않는 산출물은 이용·생성되었다 하더라도 저작권 침해로 볼 수 없는 것이다. 그리고 AI 개발을 위한 TDM과 직접 관계된 제30조의4 제2호는 표현된 사상과 감정의 비향수적 이용형태로 ‘정보해석 용도로의 제공’을 규정한다. 정보해석은 다수의 저작물 또는 기타 대량의 정보로부터 해당 정보를 구성하는 언어, 음, 영상, 기타 요소에 관한 정보를 추출, 비교, 분류 그 밖의 분석을 실시하는 것을 의미한다. TDM은 대량의 소재에 대한 비교, 분석 등을 통해 새로운 정보를 생성하므로 정보해석의 의미와 정확히 부합한다[28].

AI 학습용 분석을 포함하여 정보 분석용으로 제공하는 경우와 같이 비향수 목적인 경우 동법 제30조의4 제2호에 해당하지만 하나의 이용행위에 비향수 목적과 향수 목적이 병존하는 경우 제30조의4가 적용될 수 없다[29]. 생성형 AI와 관련하여 두 목적이 병존한다고 평가되는 대표적인 예로는 추가 학습 중 의도적으로 학습데이터에 포함된 저작물의 창작적 표현을 그대로 출력시키는 것을 목적으로 저작물의 복제 등을 실시하는 경우가 있다. 제30조의4는 AI가 학습데이터로 저작물을 비향수적 목적으로 이용하는 경우 광범위한 TDM 면책을 인정하나 비향수적 목적에 일말의 향수 목적이 개입되면 저작권자의 이익을 침해할 소지가 있어 면책 적용이 제한된다.

한편, 일본 「저작권법」 제47조의5는 컴퓨터를 이용한 정보 처리를 통해 새로운 지식과 정보를 창출함으로써 저작물의 이용촉진에 이바지하는 행위를 하는 자가 일정한 행위에 부수하여 경미하게 저작물을 이용하는 것을 면책대상으로 규정한다. 동법 제30조의4에 따른 비향수적 이용에 해당하지 않더라도 TDM 등에 부수한 경미한 이용에 대해서는 저작권 침해가 되지 않도록 정한 것이다. 일본 「저작권법」은 AI 학습데이터 활용과 관련하여 저작자의 이익을 부당하게 해치지 않는 범위 내에서 비향수 목적의 이용에 제30조의4에 따른 TDM 면책을 인정하고 비향수적 이용에 해당하지 않더라도 학습데이터의 경미한 이용에 대하여는 제47조의5에 따른 면책을 허용한다. 이러한 체계는 다른 주요국에 비해 학습데이터 활용에 상대적으로 유연한 입장을 보여준다.

3-4 중국

2017년 중국 국무원은 “차세대 인공지능 발전 계획(新一代人工智能发展规划)”을 발표하며 중국 내 인공지능 기술 발전을 위해 3단계 계획을 수립하였다[30]. 이를 기반으로 중국 내 인공지능 기술의 발전을 확산하기 위한 체계 기반을 만들었다. 이후 중국은 온라인을 중심으로 한 산업 성장이 급격하게 증가하게 되며, 디지털 경제발전에 있어 인공지능 기술이 핵심 원동력으로 기술 혁신을 주도하는 요소로 작용되었다. 그러나 중

국 역시 AI기술이 학습하는 데이터에 대한 저작권 침해 사례들이 등장하며, 종래의 「저작권법」과 AI기술로 인한 산업적 분쟁들이 점차 나타나게 되었다.

2023년 5월 23일 중국 베이징 인터넷 법원에서 오픈소스 기반 생성형 AI와 저작권 침해 관련 사건이 접수되었다. 동 사건은 A씨는 오픈소스 기반 생성형 AI Stable Diffusion을 활용하여 제시어와 매개변수를 반복적으로 조정해 “봄바람이 전해준 상냥함” 이미지를 창작 후, SNS(Social Network Service, 이하 SNS) 플랫폼에 사건 이미지를 업로드 했다. 이후 B씨는 모 SNS 계정에 “3월의 사랑, 복숭아꽃 안에서(三月的爱情, 在桃花里)”이라는 제목으로 글을 게시하는 과정에서 A씨의 허가를 받지 않고 사건 이미지의 워터마크를 제거하여 업로드하였다. 이에 A씨는 B씨가 서명권 및 정보네트워크전파권 침해 등의 이유로 B씨를 상대로 소송을 제기했다. 해당 사건에 있어 법원은 AI가 개인한 창작물에 대해 인간의 창작적 개입이 어느 정도인지에 따라 저작권법상 저작물로 인정할 것인지 판단하였다. 이에 법원은 사건의 이미지가 생성되는 과정에서 A씨가 제시어 입력, 매개변수 설정 및 반복적 수정 과정을 통해 창작의 방향을 구체적으로 설계한 결과물로, A씨의 개성적 판단과 심미적 선택이 반영된 인간의 지적 성과물로 평가하였다. 이는 AI생성물에 있어 저작물성 판단에 창작적 개입의 정도 수준이 구체적으로 드러나는 경우 저작권으로 보호가 가능한 점을 보여주고 있다.

또한 법원은 정보네트워크전파권 침해에 대해 무단 업로드 행위가 저작권자의 전파권 침해로 인정되는지에 대해 B씨가 A씨의 허락 없이 이미지를 SNS에 업로드한 행위는 대중이 임의의 시간과 장소에서 저작물을 이용할 수 있도록 하였으므로 전파권 침해에 해당한다고 판단하였다. 또한 서명권 침해에 있어 워터마크 제거가 저작자의 성명표시권을 침해하는 행위로, 저작자의 신분을 나타내는 표지를 무단으로 제거하여 업로드한 행위는 저작자의 성명표시권을 침해한 것으로 인정하였다[31].

이처럼 디지털 환경에서 디지털 콘텐츠를 보호하기 위한 워터마크 등과 같은 저작권 보호기술은 단순한 기술을 넘어 기술적 표시가 저작자 식별 기능을 수행할 수 있음을 명확히 하고 있으며, 온라인 환경에서의 저작권 침해 사건에 대해 저작자의 동일성 식별 기준에 활용할 수 있다는 시사점을 남겼다. 2024년 2월 8일 중국 광저우 인터넷 법원은 생성형 AI 산출물에 대해 최초로 저작권 침해를 인정한 판결한 사건이 발생했다. 해당 사건은 일본 ‘울트라맨’ 시리즈의 저작권을 보유하고 있는 원고가 피고 A사의 AI서비스가 유사 이미지를 생성한다며 시작된 소송 사건이다. 해당 사건에 대해 법원은 생성된 이미지가 원작과 실질적으로 유사하다고 보며, 이에 복제권과 각색권 등을 침해하였다고 판단하였다.

각색권이란 저작물 변형을 통하여 독창성을 갖춘 새로운 저작물을 창작할 수 있는 권리로 우리나라의 2차저작물작성권에 해당한다. 또한 AI 서비스를 운영하는 피고가 생성형 인공지능 감정관리 방법(生成式人工智能服务管理暂行办法)에서의 서비스 제공자의 의무를 다하지 않았음을 이유로 배상 책임을 인정

하였다. 동 사건에서는 피고가 ‘생성형 인공지능 서비스 제공자’에 해당하고, ‘생성형 인공지능 서비스 제공자’가 책임을 면하기 위해 필요한 조치를 취하지 않았다고 보아 ‘생성형 인공지능 감정관리 방법’ 제4조, 제12조, 제15조를 위반하였다고 판단하였다. 생성형 AI 이미지가 실질적으로 원저작물과의 저작권 침해 판단 기준을 명확하게 제시하지 않았으나 생성형 인공지능 서비스 제공자가 부담해야 할 의무에 관하여 구체적으로 판단한 사건이다.

이처럼 중국 또한 AI 기술 발전에 따라 다양하게 등장하며 생성형 AI 서비스 제공자의 법적 책임 범위를 구체적으로 판시하는 등 AI서비스 제공자가 무단으로 원저작물을 무단으로 학습하지 못하도록 법제도적으로 도입하여 규제하고자 한다.

3-5 종합분석

미국은 「저작권법」의 공정이용 법리를 토대로 최근 판례에서 변형성 및 시장영향 등 요소를 종합하여 AI 학습의 적법성을 구체화하고 있으며, 우리나라도 공정이용 법리를 수용해 온 점에서 그 판례 전개를 참조할 실익이 크다. EU는 DSM 지침과 AI Act를 통해 연구기관 특례와 일반 TDM(오픈아웃 전제)을 이원화하고 범용 AI 모델에 대한 투명성·요약 의무를 부과함으로써 합법적 학습과 권리자 보호의 균형을 제도적으로 설계하고 있다.

일본은 「저작권법」 제30조의4(비향수 목적 TDM)와 제47조의5(경미한 이용)를 통해 저작자의 정당한 이익을 침해하지 않는 범위에서 학습데이터 활용을 폭넓게 허용하는 등 상대적으로 유연한 체계를 운영하고 있다.

요컨대 미국은 공정이용 법리를 판례로 구체화하고, EU와 일본은 TDM 면책과 투명성 규범을 입법으로 정립하는 추세가 병행되고 있다. 반면 우리나라는 AI 학습과 관련한 공정이용 판결이 없고 TDM 규정의 입법도 현실화되지 않아 해석상 공백이 존재한다. 그로 인해 학습데이터 활용의 법적 불확실성과 저작권자 권리침해 가능성이 동시에 증대하는 문제가 나타난다. 이러한 비교법적 관찰은 우리나라가 단기적으로는 공정이용 판단지침의 명문화 및 사례 축적을 통해 예측가능성을 높이고 중장기적으로는 TDM 면책과 권리자 오픈아웃 학습데이터 투명성 의무를 포함하는 규범 설계를 검토할 필요가 있다. 또한 중국 사례에서 살펴볼 수 있듯 AI서비스 제공자가 원저작물의 학습데이터 등을 무단으로 학습하지 못하도록 제도적으로 규제를 하고 있다. AI서비스 제공자에 대해 학습데이터에 있어 무단으로 활용되는 것을 방지하기 위해 제도로 도입하며, AI서비스 제공자의 법적 의무 체계를 두기 위한 정책적 시사점 등이 존재한다.

IV. AI 학습 불가능한 데이터 보호 방안 제언 (프레임워크 모델)

4-1 학습불가능한 예제 기술과 저작권 보호 메커니즘

본 장에서는 저작권자의 권익을 보호하고 생성형 AI의 무단

학습을 방지하기 위한 기술적 조치에 대해 논한다. 저작권자는 저작물이 불법적인 이용에 활용될 우려가 존재함에도 불구하고, 홍보 및 수익 사업을 목적으로 저작물을 인터넷에 공개한다. 다만 온라인 콘텐츠 특성상 일단 공개되면 그 사용을 통제하기 어렵다. 따라서 창작자와 데이터 소유자가 자신의 저작물을 보다 능동적이고 효과적으로 보호할 수 있는 새로운 기술적 해결책을 모색할 필요가 있다. 최근 상술한 문제를 해결하기 위한 새로운 AI 학습 방법이 연구되고 있다. 특정 학습 데이터를 AI가 학습하지 못하도록 방해하는 ‘학습불가능한 예제(Unlearnable Examples)’ 기법과 이미 학습된 AI 모델에서 특정 학습 데이터를 통해 학습된 정보만을 제거하는 ‘머신 언러닝(Machine Unlearning)’ 기법을 예로 들 수 있다. 두 기법은 각각 사전대응(Prevention) 기법과 사후대응(Response) 기법으로 분류할 수 있다.

본 논문에서는 저작권자의 저작권 보호에 중점을 맞추고 있으므로 사전대응 기법인 학습불가능한 예제 기법에 초점을 맞춰 논하려 한다. 학습불가능한 예제는 학습데이터 자체에 인간의 눈으로는 거의 감지할 수 없는 미세한 교란(perturbation)을 추가한다. 이 교란된 데이터를 무단으로 수집하여 AI 모델 학습에 사용할 경우 모델은 데이터에 담긴 유의미한 패턴이나 특징을 학습하지 못하고 그 성능이 심각하게 저하된다. 즉 그 성능이 사실상 무작위 추측(random guessing) 수준으로 떨어지게 된다. 이는 데이터를 암호화하거나 접근을 차단하는 소극적 방어에서 벗어나, 데이터의 ‘학습 유용성’ 자체를 파괴하는 적극적인 ‘데이터 포이즈닝(Data Poisoning)’의 일종으로 볼 수 있다. 학습불가능한 예제는 다양한 도메인마다 각자에 특화된 방식으로 발전하고 있다. 그 중 이미지 분류 분야는 고전적인 도메인이면서도 저작권 보호 관점에서 실용적 가치가 높아 본 연구의 핵심 대상으로 선정하였다.

1) 주요 기법 분류

이미지 분류 도메인에서 학습불가능한 예제를 생성하는 방법론은 그 기본 원리에 따라 몇 가지 주요 유형으로 나눌 수 있다. 첫 번째로 ‘오류 최소화 (Error Minimizing; 이하 EM)’ 기법이다. 일반적으로 EM 기법은 예측과 정답 사이의 차이를 줄이는 방향으로 학습함으로써 모델의 예측 성능을 높이기 위해 사용되는 AI 학습의 핵심 원리이자 일련의 방법론이다. 하지만 해당 기법을 모델의 가중치를 대상으로 수행하는 대신 타겟 데이터, 즉, 학습 데이터에 포함될 데이터를 대상으로 수행하면 데이터에 노이즈가 추가된 형태의 학습 불가능한 예제를 생성할 수 있다. EM 기법을 통해 생성된 학습 불가능한 예제는 학습 과정을 반복함에도 AI 모델이 해당 데이터로부터 더 이상 학습할 것이 없다고 판단하여 모델 업데이트가 일어나지 않는다[32].

EM 기법은 활용 목적에 따라 각 데이터 샘플마다 특화된 노이즈를 생성하는 sample-wise 전략과 같은 클래스에 속하는 모든 샘플에 동일한 노이즈를 생성하여 class-wise 전략으로 나뉜다. 그러나 초기의 EM 기법을 통해 생성된 학습 불가능한 예제는 적대적 학습(Adversarial Training) 등 강건성

(Robustness)이 높은 모델에서 보호 효과가 감소하는 한계가 존재했다. 이러한 한계를 보완하기 위해 SEM(Stable Error Minimization, 이하 SEM) 등의 기법이 제안됐다. SEM은 기존 EM 방식의 강건성을 개선하기 위한 기법으로, EM이 적대적 교란을 기반으로 노이즈를 최적화하는 것과 달리, 무작위 교란을 기반으로 노이즈를 학습시킨다. 이로 인해 SEM은 방어 노이즈의 안정성과 일반화 능력이 향상되어, 적대적 학습과 같이 강건하게 훈련된 모델을 상대로도 학습 불가능한 특성을 잘 유지하는 것으로 알려져 있다.

두 번째로 ‘합성곱 기반(Convolution-based)’ 접근법은 주파수 영역에서 클래스별 패턴을 삽입하여 데이터를 학습 불가능하게 만든다. 대표적인 CUDA(Convolution-based Unlearnable Datasets) 기법은 각 클래스마다 고유한 합성곱 커널을 설계하여 이미지에 적용한다[33]. 예를 들어, 고양이 이미지들에는 특정 주파수 패턴 A를, 개 이미지들에는 패턴 B를 삽입하는 방식이다. AI 모델은 실제 동물의 특징 대신 이러한 인위적 패턴만을 학습하게 되어, 새로운 이미지를 만났을 때 전혀 분류하지 못하게 된다. 이 방법은 기존의 복잡한 최적화 과정 없이도 빠르게 보호된 데이터를 생성할 수 있어 실용적이며, 주파수 영역에서 작동하기 때문에 시각적으로는 원본과 거의 구별되지 않는다.

세 번째로 ‘특징 기반(Feature-guided) 접근법’은 AI 모델의 내부 특징 공간을 직접 조작하여 학습을 방해한다. 이 방법들은 같은 클래스의 데이터들이 서로 멀리 떨어지도록 하거나, 반대로 다른 클래스의 데이터들이 뒤섞이도록 유도한다. 예를 들어, TUE(Transferable Unlearnable Examples) 기법은 고양이 이미지들의 특징을 서로 흩어지게 만들면서 동시에 개나 새 이미지의 특징과 섞이도록 조작한다[34]. EntF(Entangled Features) 기법은 더 나아가 서로 다른 클래스의 샘플들이 공통된 ‘얽힌 특징’을 갖도록 만들어, AI가 어떤 특징이 어떤 클래스에 속하는지 구분할 수 없게 만든다[35]. 이러한 접근법은 단순히 노이즈를 추가하는 것이 아니라 AI의 학습 메커니즘 자체를 교란시켜 더욱 강력한 보호 효과를 달성한다.

2) 상용 어플리케이션 분석

Glaze는 시카고 대학교 연구팀이 개발한 도구로, 예술가들의 작품을 텍스트-이미지 생성 모델(예: Stable Diffusion, Midjourney)로부터 보호하기 위해 설계되었다[36]. Glaze는 스타일 모방(style mimicry)을 방지하는 데 특화되어 있으며, 작품에 시각적으로 감지하기 어려운 교란을 추가하여 AI 모델이 특정 예술가의 스타일을 학습하지 못하도록 한다. 실험 결과, Glaze로 보호된 작품으로 학습한 모델은 원본 스타일을 재현하는 능력이 현저히 저하되었다. Nightshade는 같은 연구팀이 개발한 더 공격적인 도구로, 단순히 학습을 방해하는 것을 넘어 AI 모델을 ‘오염’시킨다[37]. 예를 들어, 개 이미지를 고양이로 인식하도록 하는 잘못된 교란을 삽입하여, 이를 학습한 모델이 완전히 잘못된 출력을 생성하도록 유도한다. Nightshade는 소수의 오염된 샘플만으로도 전체 모델 성능을 크게 저하시킬

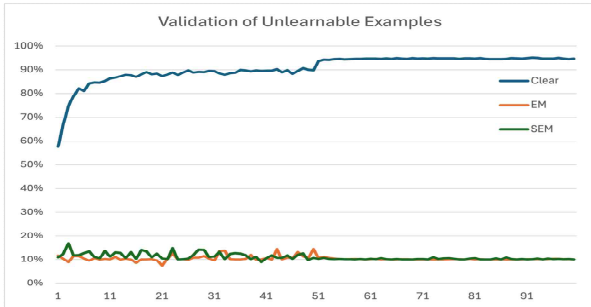


그림 1. 학습 불가능한 예제 생성 시 정확도 하락 검증 결과
Fig. 1. Verification results of accuracy degradation when generating unlearnable examples

수 있어 강력한 억제 효과를 갖는다.

3) 이미지 생성형 AI 대상 학습 불가능한 예제

최신 연구에 따르면, 학습 불가능한 예제 기술은 전통적인 이미지 분류 태스크를 넘어 생성형 AI, 특히 확산 모델(Diffusion Models)에 대해서도 강력한 보호 효과를 나타내고 있다. 이는 DALL-E 2, Stable Diffusion, Midjourney와 같은 텍스트-이미지 생성 모델의 급속한 발전과 광범위한 상용화 맥락에서 특히 중요한 의미를 갖는다. 보호된 데이터로 훈련 시 FID(Frechet Inception Distance) 점수가 50~100% 증가한다. FID는 생성된 이미지와 실제 이미지 분포 간의 거리를 측정하는 지표로 낮을수록 생성 품질이 높다. 학습 불가능한 예제로 보호된 데이터로 훈련 시 FID 점수가 50~100% 증가한다는 것은 생성된 이미지가 실제 이미지 분포에서 크게 벗어났음을 의미한다. 또한 CLIP 유사도 점수가 40~60% 감소하는데 CLIP Score는 생성된 이미지와 텍스트 프롬프트 간의 의미적 일치도를 측정하므로 생성형 AI 시대에 학습 불가능한 예제가 실질적 보호 수단임을 증명한다.

4-2 학습 불가능한 예제 기술의 실효성 검증

본 연구는 이미지 분류 도메인에서 학습 불가능한 예제의 실효성을 검증하기 위해 일련의 실험을 수행하였다. 실험은 표 1의 환경에서 ① 보호되지 않은 원본 데이터를 이용한 학습, ② 학습 불가능한 예제가 적용된 데이터를 이용한 학습, ③ 적대적 학습(Adversarial Training) 환경에서의 견고성 평가의 세

표 1. 실험 환경 구성 정보

Table 1. Experimental environment configuration information

Dataset	CIFAR-10
UE Algorithm	EM, SEM
Metrics	Accuracy
Hardware	CPU: AMD Ryzen 3950x GPU: NVIDIA GeForce RTX 4090
Software	Cuda 12.2 Python 3.10.12 Pytorch 1.8.1 Torchvision 0.9.1

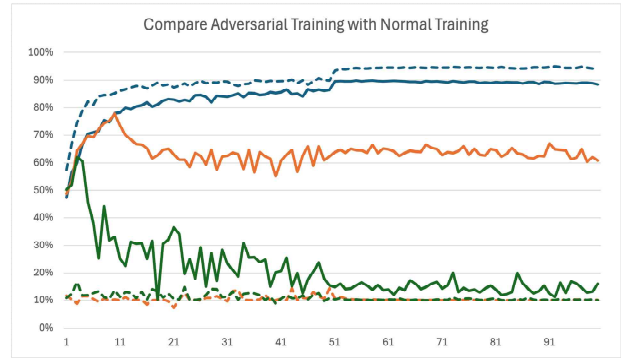


그림 2. 적대적 학습에도 강건한 학습 불가능한 예제의 성능 검증 결과
Fig. 2. Verification results of the performance of unlearnable examples robust to adversarial training

가지 시나리오를 기반으로 진행되었다. 먼저 그림 1에서 볼 수 있듯이 기준 모델(ResNet-18)의 원본 데이터 학습 시 분류 정확도는 94.95%에 달했으나, EM 기법이 적용된 데이터로 학습한 경우 정확도가 10.09%(학습 종료 시점 정확도)로 급격히 저하되었다. 이는 무작위 추측(10%)과 유사한 수준으로, 학습 불가능한 예제가 AI 모델의 학습을 효과적으로 방해함을 보여준다. SEM 기법은 10.05%(학습 종료 시점 정확도)의 정확도를 기록하여 모든 기법이 유의미한 보호 효과를 나타냈다. 그림 2는 동 데이터셋 및 동 기법에 대해 적대적 학습을 수행한 결과를 보여준다. 이 그래프에서 점선은 일반학습 시의 결과이며, 실선은 적대적 학습 시의 결과를 나타낸다. 여기서의 적대적 학습은 일반적인 방식으로 모델을 훈련시킨 것과 달리, 훈련 과정에서 PGD(Projected Gradient Descent) 공격으로 생성된 교란된 입력을 모델에 함께 학습시켜 의도적으로 모델의 강건성을 높이는 방식이다. 본 연구는 이처럼 강건하게 훈련된 모델이 EM과 SEM으로 보호된 데이터의 학습 불가능한 특성을 무력화시키는지, 즉 보호 효과가 적대적 학습 환경에서도 유지되는지를 확인하고자 하였다. 해당 결과를 통해 EM 기법은 적대적 학습 시 약 50.68%의 정확도 상승을 보였지만 여전히 원본 학습데이터로 학습했을 때의 정확도 88.43%보다 현저히 낮은 정확도를 유지한 것을 확인할 수 있다. 뿐만 아니라 SEM 기법의 경우 적대적 학습에도 약 6.06%의 정확도 상승만을 보이며 강력한 학습 불가능한 예제의 생성이 가능함을 보였다. 이는 실제 배포 환경에서의 실용성을 시사한다. 실험 결과는 학습 불가능한 예제 기술이 「저작권법」 제104조의3에 규정된 「기술적 보호조치」로서의 요건을 충족할 가능성을 보여준다. 「저작권법」상 기술적 보호조치는 저작권 등의 침해해 효과적으로 방지하거나 억제하는 기술적 조치로, 학습 불가능한 예제와 같이 AI의 무단 학습을 억제하는 기술적 장치로 기능할 수 있을 것이다. 나아가 학습 불가능한 예제는 원천 데이터에 기술적 조치를 통해 AI모델의 특정 패턴 학습 능력을 기술적으로 저해하고, 학습 결과로 생성된 모델의 유용성을 감소시키게 된다. 이는 저작권 등의 침해 방지의 실질적 효과를 가지

고 올 것으로 기대된다. 그러나 법적 실효성 측면에서 ‘효과적’이라는 요건의 해석에 있어 법원의 판단이 필요하다. 종래 우리 법원의 태도에 따르면, 저작권 보호를 위한 기술적 보호조치에 대해 단순히 접근 통제만을 목적으로 하는 조치는 기술적 보호조치에 해당하지 않으며, 저작권 침해를 효과적으로 방지하거나 억제하는 조치에 한정하여 그 법적 효력을 인정하는 태도를 보여왔다(대법원 2006. 2. 24. 선고 2004도2743 판결). 이처럼 우회 기술의 지속적 발전에 따른 기술적 군비경쟁 양상도 고려할 필요가 있다.

4-3 저작권 보호를 위한 기술적 보호조치 프레임워크

1) 이중 주체 기반 보호 메커니즘

본 연구에서 제안하는 기술적 보호조치 프레임워크는 콘텐츠 창작자(Contents Creator)와 AI 모델 학습자(AI Model Trainer) 간의 상호작용을 중심으로 설계된다. 이는 기존의 일방적 보호조치와 달리, 양 당사자 간의 권한 확인과 기술적 협조를 통한 체계적 보호 시스템이다.

2) 콘텐츠 창작자 측 프로세스

콘텐츠 창작자는 저작권 콘텐츠 x 를 보유한 상태에서 보호 필요성을 판단한다. 보호가 필요하다고 판단될 경우, 학습 불가능한 예제 노이즈 δ 를 생성하여 원본 콘텐츠에 추가한다($x_{ue} = x + \delta$). 이렇게 생성된 보호 콘텐츠 x_{ue} 는 시각적으로는 원본과 거의 구별되지 않으나 AI 학습에는 유해한 교란을 포함하게 된다. 보호조치가 불필요하다고 판단되는 경우에는 원본 콘텐츠를 그대로 게시한다. 이 과정에서 중요한 것은 노이즈 생성 알고리즘의 선택과 교란 강도의 설정이다. 저작권자는 콘텐츠의 성격, 예상되는 위협 수준, 시각적 품질 요구사항 등을 종합적으로 고려하여 최적의 보호 매개변수를 결정해야 한다.

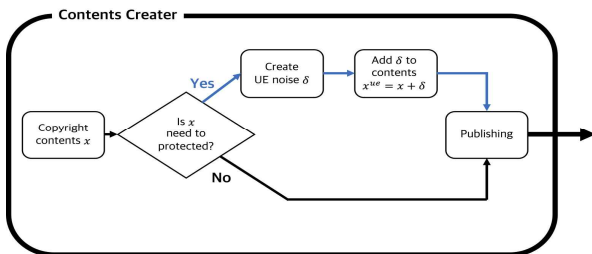


그림 3. 저작권자 관점에서의 학습 불가능한 예제 생성 프레임워크
Fig. 3. Framework for generating unlearnable examples from the copyright holder's perspective

3) AI 모델 학습자 측 프로세스(그림 4)

AI 모델 학습자는 콘텐츠 수집 단계에서 수집된 데이터의 보호 여부를 확인해야 한다. 수집된 데이터가 학습 불가능한 예제로 보호되어 있지 않은 경우, 일반적인 학습 프로세스를 통해 성공적으로 모델을 훈련시킬 수 있다.

그러나 수집된 데이터가 보호되어 있을 경우, AI 모델 학습

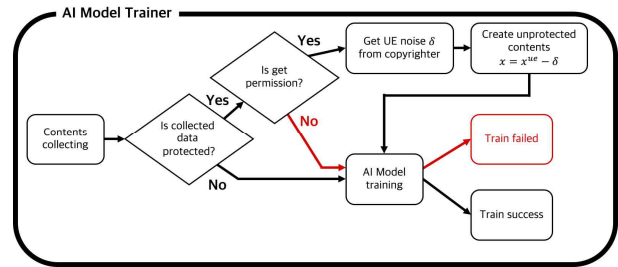


그림 4. AI 모델 생성자 관점에서의 학습 불가능한 예제 활용 프레임워크

Fig. 4. Framework for utilizing unlearnable examples from the AI model developer's perspective

자는 저작권자로부터 적법한 이용 허가를 획득해야 한다. 허가를 받은 경우, 저작권자로부터 해당 노이즈 δ 에 대한 정보를 제공받아 보호된 콘텐츠로부터 원본을 복원할 수 있다($x = x_{ue} - \delta$). 이는 정당한 이용자만이 콘텐츠의 학습 가치를 온전히 활용할 수 있도록 하는 선택적 접근 통제 메커니즘이다. 반면 적법한 허가 없이 보호된 데이터로 학습을 시도할 경우, AI 모델의 성능은 심각하게 저하되어 학습 실패로 귀결된다. 이는 무단 이용에 대한 기술적 제재 효과를 나타낸다.

4) 옵트아웃(Opt-out) 시스템과의 연계

EU의 DSM 지침과 AI Act에서 규정하는 권리자 옵트아웃 시스템과 학습 불가능한 예제 기술을 연계하는 방안을 모색해야 한다. 권리자가 robots.txt 파일이나 기계판독가능한 메타데이터를 통해 AI 학습 거부 의사를 표명한 경우, 자동으로 해당 콘텐츠에 학습 불가능한 예제 기술이 적용되는 통합 시스템을 구축할 수 있다.

이러한 시스템은 저작권자의 의사 표명을 기술적으로 강제할 수 있는 수단을 제공하여, 단순한 선언적 거부를 넘어 실질적인 보호 효과를 달성할 수 있다.

4-4 「저작권법」상 AI 서비스 제공자에 대한 의무 확대 방안

생성형 AI 시대, 저작물 등을 포함한 학습데이터에 대한 침해가능성이 현저히 높아지면서 저작권자의 입장에서 기술적 보호 조치를 보다 필요로 할 수 밖에 없는 국면에 이르렀다. 저작권자는 AI 서비스 제공자가 AI 모델에 본인의 저작물과 관계된 데이터를 학습시키는 것을 확인할 수 있는 방법이 사실상 AI가 생성한 표현을 확인하는 것 밖에 없어 허락받지 않는 저작물을 생성형 AI가 기계적으로 학습할 경우 기계 학습에 사용된 데이터를 확인할 수 없다는 점에서 권리자측에서는 침해를 입증하기 어렵다는 것이다. 이에 기술적 보호조치의 한 기법인 학습 불가능한 예제는 학습데이터 자체에 인간의 눈으로는 감지하기 어려운 미세한 교란을 추가하여 AI 모델의 학습을 방해하는 것으로 저작권자는 이를 활용하여 본인의 저작물의 심미적인 미각 효과는 유지하면서도 AI의 무차별적인 학습으로부터 보호받을 수 있다. 다만 저작권을 보호하고자 저작물에 기술적 보호조치를 적용하는 경우 기술조치의 효과는 법과는 관

제없이 이를 적용하는 사람의 의도에 따라 또는 그 의도에 관계없이 저작권의 경계를 넘어서게 된다[38]. 헌법재판소 2018. 11. 29. 선고 2017헌바369 전원재판부 결정의 태도에서 살펴볼 수 있듯 저작물의 복제·전송 등을 통제하기 위한 기술적 수단을 의미하며, 저작권 등 권리의 침해 행위 자체를 방지 또는 억제하는 기술적 보호조치와 그와 동등한 효과가 있는 기술적 보호조치가 모두 포함된다고 해석되고 있다. 따라서 기술적 보호조치는 그 기본적인 속성상 저작권이 부여된 저작물의 불법복제나 유통을 막기 위한 것이나 필요 이상으로 저작물의 이용과 유통을 통제할 수 있기 때문에 「저작권법」이 일정한 경우 공익을 위하여 저작권을 제한하거나 자유이용을 허용하는 경우에 있어서 조차 저작물의 이용이 불가능하게 될 수 있다[39].

이에 따라 기술적 보호조치는 저작물의 유통의 관점에서 봤을 때 저작권자가 아닌 AI 서비스 제공자에게 그 의무를 부담하게 하는 것이 합리적이라 할 수 있다. 또한 TDM 면책 규정과 「저작권법」상 공정이용의 법리를 비추어 봤을 때에도 AI 서비스 제공자에게 기술적 보호조치에 관한 의무를 지우는 것이 합당하다. 먼저 TDM 면책 규정이 있는 EU와 일본의 경우 저작권자의 이익을 침해하지 않는 범위 내에서 권리자의 옵트아웃 또는 비향수적인 목적으로 저작물을 이용하면 TDM 면책 규정을 허용하므로 AI 서비스 제공자의 법적인 부담이 상대적으로 덜하다. 반면 미국과 우리나라는 TDM 면책 규정이 아닌 「저작권법」상 공정이용의 법리를 활용하는데 최근 미국에서 AI의 개발을 위해 저작물을 학습시키는 행위가 공정이용에 해당하는지 여부에 관한 판결이 선고되며 이러한 미국의 판결이 우리나라에서 중요한 참고자료가 될 것으로 보는 시각도 존재한다[21].

이는 우리나라도 향후 공정이용의 법리 제1요소부터 제4요소까지 종합적으로 고려하여 AI의 저작물 학습의 적법성을 두고 논의가 이루어질 가능성이 높다는 점에서 AI 서비스 제공자는 불확실한 제도적 기반에 따라 부담을 느낄 것으로 생각한다. 「저작권법」의 공정이용은 AI 서비스 제공자가 AI 개발을 위해 저작물 등의 학습데이터 활용의 정당성을 주장하기 위해 일반적으로 사용되나 저작권자는 자신의 이익을 부당하게 침해한다고 여겨 법원에서의 공방 기간이 발생할 수 밖에 없다. 해당 기간에 생성형 AI가 만든 창작물은 제3자가 활용할 수 있기 때문에 저작권자의 이익을 침해하는 상황이 발생할 수 있어 해당 창작물 및 이를 생성하는데 활용될 수 있는 학습데이터에 대한 AI 서비스 제공자의 기술적 보호조치가 의무화되어야 한다. 우리 「저작권법」은 제2조에서 기술적 보호조치를 ① 저작권, 그 밖에 이 법에 따라 보호되는 권리의 행사와 관련하여 이 법에 따라 보호되는 저작물 등에 대한 접근을 효과적으로 방지하거나 억제하기 위하여 그 권리자나 권리자의 동의를 받은 자가 적용하는 기술적 조치 ② 저작권, 그 밖에 이 법에 따라 보호되는 권리에 대한 침해 행위를 효과적으로 방지하거나 억제하기 위하여 그 권리자나 권리자의 동의를 받은 자가 적용하는 기술적 조치로 정의한다. 동법에서 기술적 보호조치를 시행해야 하는 대표적인 주체로는 온라인 서비스 제공자가 있다. 「저작권

법」 제104조에서 온라인 서비스 제공자는 권리자의 요청이 있는 경우 해당 저작물 등의 불법적인 전송을 차단하는 기술적인 조치 등 필요한 조치를 하여야 한다. 해당 조항의 대상 범위를 AI 서비스 제공자도 포함하여 저작권자가 본인의 저작물이 불필요하게 유통되는 것을 방지할 수 있게 AI 서비스 제공자에게 의사를 표시할 경우 기술적 보호조치를 시행하도록 하여 저작권자의 권리를 보호할 필요가 있다. 또한 최근에는 지브리 등 스타일이 누가 봐도 공중의 영역이 아닌 지브리풍으로 인식되는 사안에 관해서도 「저작권법」의 아이디어·표현 이분법 원칙에 따라 AI의 학습을 저작권 침해로 보지 않으면서 저작권자의 창작 의욕을 심각하게 저하시킬 수 있는 상황에 직면해 있다. 이러한 법적 공백을 해소하기 위해 기술적 보호조치가 필요하므로 「저작권법」에 AI 서비스 제공자에 대한 의무를 확대할 수 있도록 해야 한다.

한편, 2024년 12월 16일 국회 본회의를 통과하여 공포 후 1년 뒤에 시행될 예정인 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 “인공지능기본법”）」은 EU AI법과 비교·분석했을 때 규제보다는 산업 진흥에 초점이 맞춰져 있어 AI 서비스 제공자는 해당 법률을 기반으로 저작물 등이 포함된 학습데이터를 활용할 수 있다. 이에 따라 「저작권법」과 「인공지능기본법」의 충돌이 일어날 수 있는 가운데 「저작권법」이 저작권자와 AI 서비스 제공자의 균형있는 이익을 모색하고 문화 및 산업발전에 이바지하는 법률로 나아가기 위해서는 시대의 변화에 맞춰 기술적 보호 조치 대상의 범위를 온라인 서비스 제공자뿐 아니라 AI 서비스 제공자로 확대하여 동법이 지니는 법적 공백을 해소해야 할 것으로 보인다.

V. 결 론

본 논문은 생성형 AI 시대, AI 시스템의 핵심 원료인 학습데이터를 활용하는 과정에서 발생될 수 있는 법적 문제와 제도적 한계점 등을 종합적으로 분석하였다. 또한 주요국들에서의 AI시스템 내 학습데이터의 활용에 있어 제도적 동향을 분석하였다. 이를 통해 전통적인 제도적 한계 등으로 인해 발생하는 AI시스템이 학습데이터를 활용하며 발생하는 법적 쟁점들을 중심으로 제도적·기술적 관점에서 교차 분석하고, 이를 해결하기 위한 새로운 방식의 프레임워크를 제시하였다. 본 연구는 선행 연구와 달리 다음과 같은 핵심적 기여는 다음과 같이 제시할 수 있다.

먼저 국내 공정이용 해석의 불확실성과 아이디어·표현 이분법 경계의 문제점을 제시하였다. 이에 생성형 AI 시스템이 공개된 데이터 등을 수집하여 훈련하는 단계에서 발생하는 학습행위들이 적법성에 관한 예측가능성이 현저히 낮다는 점을 확인하였다. 주요국들의 제도적 분석을 위한 비교법적 연구를 통해 미국의 경우 공정이용에 대해 4요소를 종합적으로 판단하는 판례 경향을 통해 ‘변형성’과 ‘시장 영향’을 중심으로 사실상 규범적 가이드라인으로 축적하고 있었다. 반면 EU와 일본의

경우 TDM 면책과 투명성 의무를 제도화함으로써 합법적 학습데이터에 대한 경로를 명시적으로 제공하고 있다. 반면 우리나라의 경우 공정이용에 대한 판례 및 명확한 기준이 모호하여, 생성형 AI 시스템이 학습데이터를 활용 및 법적 안정성 등에 대한 공백 상태에 머물러 있음을 도출하였다. 이는 AI산업은 물론 저작권 관련 산업에서의 눈에 보이지 않는 규제로 작용될 수 있다.

둘째, 화풍은 원칙적으로 아이디어의 영역으로 저작권 보호의 직접적인 대상이 되기 어렵다. 그러나 최근 ‘지브리풍’ 논쟁 등으로 인해 저작권 이슈 보다 부정경쟁 등 타 제도적 규율로 우회될 수 있다. 그러나 실제 산업 현장에서는 스타일의 특정 출처와 함께 강하게 결부되는 경계 사례들이 증가하고 있어 표현 보호가 사실상 아이디어 보호로 전이될 수 있다는 쟁점들이 존재한다. 이는 저작권 보호 및 활성화 등의 체계가 창작자의 인센티브를 유지하며, 공중 영역을 과도하게 잠식하지 않도록 하는 균형적인 제도적 기반 마련이 필요하다.

마지막으로 현행법 체계에서의 제도적 공백을 보완하기 위해 생성형 AI 등에 특화된 기술적 보호조치의 실효성을 제시하였다. 이를 위해 이미지 분류 도메인에서 학습 불가능한 예제 기법을 적용하면 AI 시스템의 정확도가 사실상 무작위 추측 수준으로 붕괴함을 확인하였다. 이는 데이터 접근 또는 열람이 아닌 학습 유용성 자체를 파괴하는 적극적 보호 기법이라는 점에서 저작권법상 기술적 보호조치 개념에 부합한다. 따라서 생성형 AI 등에 있어 무단으로 학습되는 데이터에 대해 사전적 억제수단으로 작용되기 위해 생성형 AI 시스템에 대해 저작권법상 기술적보호조치의 적용이 필요하다는 점을 제시하였다.

그러나 본 연구에도 한계점은 다음과 같다. 첫째 학습 불가능한 예제는 본질적으로 적대적 공간화에 직면하고 있다. 또한 우회·정화 기술의 지속적 고도화에 대응하기 위해서는 적응적 방어 전략과 강건성 평가 프로토콜의 정립이 요구된다. 둘째 텍스트·오디오·비디오 등 도메인 및 매체 간 전이 가능성에 대한 체계적 분석과 시각적 품질과 보호 강도 간의 불가피한 상충관계를 정량적으로 규명하는 연구가 필요하다. 그럼에도 불구하고 본 연구는 주요국들의 생성형 AI 시스템이 활용하는 학습데이터에 있어 제도적 한계와 공백을 분석하여 이에 대한 기술적 보호조치를 위한 기술적 제언 등을 제시하며, AI서비스 사업자에 대해 기술적 보호조치 의무를 두어 AI 산업 발전은 물론 학습데이터의 권리자의 권리 보호를 위한 제도적·기술적 제언을 제시하였다. 본 연구에서 제언한 프레임워크는 권리자와 AI서비스 사업자, 이용자의 정당한 활용과 이익 등을 균형적으로 조화시키며, AI서비스 사업자는 학습데이터의 책임 있는 활용과 지속 가능한 AI산업 발전을 위한 산업적 법적 공백을 해소하기 위한 첫 발걸음이 될 것으로 기대한다. 향후 해당 연구를 기반으로 이미지, 영상 등 다양한 학습데이터의 유형에 맞춰 기술적 보호조치를 위한 실험적 검증 등을 통해 정책적으로 내재화할 수 있도록 제도와 기술을 융합한 연구들이 지속적으로 수행될 필요가 있다.

감사의 글

본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-정보통신방송혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2025-RS-2023-00266605, 40%).; 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 문화체육관광 연구개발사업으로 수행되었음(과제명: 블록체인 기술 기반 SW 저작권 보호를 위한 유통·관리 플랫폼 기술 개발 및 인재양성, 과제번호: RS-2023-00228867, 30%).; 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재4.0 사업의 연구결과로 수행되었음(IITP-2022-RS-2022-00156310, 30%).

참고문헌

- [1] J.-H. Park, B. Kang, C.-N. Lee, W.-Y. Yoo, and H.-K. Koh, A Guide to Copyright and Generative AI, Korea Copyright Commission, Seoul, 2023.
- [2] H. W. Cha, Copyright Law in the Age of Generative AI: Focusing on the Question of Authorship, Master's Thesis, Seoul National University, Seoul, 2025.
- [3] S. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. London, UK: Pearson, 2021.
- [4] Exploding Topics. Number of Parameters in GPT-4 (Latest Data) [Internet]. Available: <https://explodingtopics.com/blog/gpt-parameters>.
- [5] H. K. Park, "A Study on the Use of Copyrightable Works in Machine Learning", *The Journal of Sports and Entertainment Law*, Vol. 23, No. 1, pp. 129-152, 2020. <http://dx-doi-org.proxy.cau.ac.kr/10.19051/kasel.2020.23.1.129>
- [6] G. J. Yoon, The Limitations of Fair Use of Works for Generative AI Training, Korea Legislation Research Institute, Sejong, A Issue Paper, 2023.
- [7] T. B. Kwon, "Idea Protection System in Intellectual Property Right", *Journal of Industrial Property*, No. 41, pp. 1-41, 2013. <https://www-earticle-net.proxy.cau.ac.kr/Article/A255004>
- [8] S. J. Oh, *The Copyright Law*, 4th ed. Seoul: Pakyoungsa, 2023.
- [9] D. H. Lee and H. S. Roh, "Rethinking the Idea-expression Dichotomy", *Journal of Business Administration & Law*, Vol. 23, No. 2, pp. 617-657, 2013.
- [10] S. H. Park, *The Copyright Law*, 3th ed. Seoul: Pakyoungsa, 2023.
- [11] Seoul High Court Decision, 2015Na2063761, rendered on January 12, 2017.
- [12] The Chosun Ilbo. Ghibli-style AI Images Could Be Illegal

- in Korea [Internet]. Available: https://www.chosun.com/economy/tech_it/2025/04/09/K522W6ZIK5GJBFRWVTY36LHSU4/.
- [13] M. J. Kim, “AI That Creates Ghibli-style Images: Any Copyright Issues?,” *The Journal of Monthly Secu N*, No. 196, pp. 40-43, May 2025.
- [14] S. J. Choi, “Artistic Works in the Era of Artificial Intelligence in the Context of the Copyright Law and Unfair Competition Prevention Act,” *The Journal of Korea Information Law*, Vol. 29, No. 2, pp. 87-116, 2025.
- [15] J. W. Park, “Can ‘Selection or Arrangement’ Alone Qualify an AI-Generated Work as a Compilation? - Focusing on Zarya of the Dawn,” *The Journal of IP& Data Law*, Vol. 5, No. 1, pp. 84-86, 2025.
- [16] S. J. Choi, Copyright Infringement Issues of AI-Generated Works Using Specific Authors’ Styles as Prompts, Korea Copyright Commission, Seoul, Copyright Issue Report, 2025.
- [17] *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522-23 (9th Cir. 1992); *American Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 922 (2d Cir. 1994).
- [18] Y. S. Park, “Transformative Use Test in the U.S. Copyright Law,” *The Journal of Copyright Quarterly*, Vol. 28, No. 2, pp. 73-99, 2015.
- [19] SHIN & KIM. Recent U.S. Case Law Trends on Fair Use and the Use of Copyrighted Works for AI [Internet]. Available: <https://www.lawtimes.co.kr/LawFirm-NewsLetter/209694>.
- [20] *Williams & Wilkins Co. v. United States*, 487 F.2d 1345, 1353 (Ct. Cl. 1973), *aff’d* by an equally divided Court, 420 U.S. 376 (1975); *Sony Corp. of America v. Universal City Studios*, 464 U.S. 417, 449-50 (1984); *Hustler Magazine v. Moral Majority*, 796 F.2d 11.
- [21] The Law Times. Recent U.S. Case Law Trends on Fair Use and the Use of Copyrighted Works for AI [Internet]. Available: <https://www.lawtimes.co.kr/LawFirm-NewsLetter/209694>.
- [22] AI TIMES. Anthropic Pays 2 Trillion Won in Settlement First Case Recognizing AI Copyright [Internet]. Available: <https://www.aitimes.com/news/articleView.html?idxno=202163>.
- [23] J. H. Hong, U.S. Court’s Fair Use Judgment on Meta’s Use of Works for AI Training, Korea Copyright Commission, Seoul, Copyright Trend Report No. 7, 2024.
- [24] J. H. Park and D. H. Park, TDM Regulations in the European Union and Japan and Their Scope of Application, Korea Copyright Commission, Seoul, Copyright Trend Report No. 7, 2024.
- [25] Personal Information Protection Commission and European Commission. Guidelines on the Scope of Obligations for General-Purpose AI (GPAI) Models [Internet]. Available: www.pipc.go.kr
- [26] The Law Times. Recent Legal Issues Concerning the Training and Development of AI Models [Internet]. Available: <https://www.lawtimes.co.kr/LawFirm-NewsLetter/206962>.
- [27] N. Nakayama, *Copyright Law*, 3rd ed. Yuhikaku, 2020.
- [28] S. Ryu, “A Study on Introducing a Specific Copyright Exception Clause to Text and Data Mining,” *The Journal of Advanced Commercial Law Review*, No. 101, pp. 347-390, 2023.
- [29] Y. Y. Jang, Report on Japan Cultural Affairs Agency’s Publication on AI and Copyright, Korea Copyright Commission, Seoul, Copyright Issue Report, 2024.
- [30] National Health Commission of the People’s Republic of China. Notice of the State Council on Issuing the New Generation Artificial Intelligence Development Plan [Internet]. Available: <https://www.nhc.gov.cn/bgt/gwywj2/201707/3b2a93a71c794d9c8137ab394b21d8f3.shtml>.
- [31] Beijing Court, Civil Judgment 2023 Jing 0491 Minchu No. 11279.
- [32] Y. Liu, K. Xu, X. Chen, and L. Sun, “Stable Unlearnable Example: Enhancing the Robustness of Unlearnable Examples via Stable Error-Minimizing Noise,” *arXiv:2311.13091*, 2024. <https://doi.org/10.48550/arXiv.2311.13091>
- [33] V. S. Sadasivan, M. Soltanolkotabi, and S. Feizi, “Cuda: Convolution-Based Unlearnable Datasets,” *arXiv:2303.04278*, 2023. <https://doi.org/10.48550/arXiv.2303.04278>
- [34] J. Ren, H. Xu, Y. Wan, X. Ma, L. Sun, and J. Tang, “Transferable Unlearnable Examples,” *arXiv:2210.10114* 2023. <https://doi.org/10.48550/arXiv.2210.10114>
- [35] R. Wen, Z. Zhao, Z. Liu, M. Backes, T. Wang, and Y. Zhang, “Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning?” in *The Eleventh International Conference on Learning Representations*, 2023.
- [36] GLAZE. Main Page [Internet]. Available: <https://glaze.cs.uchicago.edu>.
- [37] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, “Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models,” in *Proceeding of the 2024 IEEE Symposium on Security and Privacy (SP)*, San Francisco: CA, pp. 807-825, 2024. <https://doi.org/10.48550/arXiv.2310.13828>
- [38] P. Samuelson, “DRM(and, or, vs.) the Law,”

Communications of the ACM, Vol.46, No.4, pp. 41-45, April 2003. <https://doi.org/10.1145/641205.641229>

- [39] H. S. Tak, "A Study on the Circumvention against Technological Protection Measures for Copyrighted Works", *Journal of Korean Criminological Review*, Vol. 20, No. 1, pp. 1229-1250, March 2009.



김종성(Jongseong Kim)

2021년 : 중앙대학교 융합보안학 석사
2024년 : 중앙대학교 보안경영학 박사

2020년~2025년: 한국저작권보호원 주임

2025년~현 재: 중앙대학교 융합보안학과 연구교수

※관심분야: 정보보호, 디지털저작권, 생성형 AI, 산업보안법 등

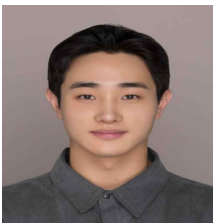


박성환(Sunghwan Park)

2021년 : 중앙대학교 융합보안학 석사

2021년~현 재: 중앙대학교 일반대학원 융합보안학과 박사수료

※관심분야: 인공지능 보안, 연합학습, 분산시스템 보안 등



황요한(Yohan Hwang)

2023년 : 경기대학교 지식재산학과
경제학사

2024년~현 재: 중앙대학교 융합보안학과 석사과정

※관심분야: 저작권법, 산업보안법, 디지털포렌식 등



오병훈(Byunghoon Oh)

2025년 : 명지대학교 수학과 (학사)

2025년~현 재: 중앙대학교 융합보안학과 석사과정

※관심분야: 인공지능 보안, 적대적 학습, 데이터 분석



이재우(Jaewoo Lee)

2008년 : 서울대학교 공학석사

2017년 : University of Pennsylvania
Computer and Information
Science 공학박사

2018년~현 재: 중앙대학교 산업보안학과 부교수

※관심분야: Real-time system, CPSsecurity 등



홍준호(Junho Hong)

2014년 : 단국대학교 법학과 법학석사

2018년 : 단국대학교 법학과 법학박사

2014년~2024년: 한국정보보호산업협회 한국정보보호교육원
원장

2024년~현 재: 성신여자대학교 융합보안공학과 조교수

※관심분야: 정보보호, 디지털저작권, 개인정보보호 등