

미국

미국에서의 AI 저작권 소송(2)

고려대학교 법학전문대학원/교수
이대희

현재 미국에서는 AI 저작권 침해 관련 소송이 66건이 제기되어 있는 상태이다.(2025년 12월 14일 기준) 이들 소송은 AI 학습을 위한 데이터 수집, AI 산출물 생성, 검색증강생성(RAG, retrieval-augmented generation) 등을 중심으로 복제권, 배포권, 전시권, 2차적저작물작성권 침해를 대상으로 하고 있다. 소송의 원고들은 언론사나 문학 작가 등 어문저작물, 사진저작물, 음악저작물 등에 대한 저작권자이고, 피고들은 대규모 AI 개발사를 망라하고 있다. 이 글은 제기된 소송들에 대하여 계류 중인 법원, 원고가 주장하는 사실관계 및 위반 사항을 중심으로 간단하게 정리한 글이다. 이 글은 66건의 소송 중 앞서 다룬 20개의 소송 외에 추가로 15개를 다루고 있고, 나머지 소송들은 동일한 제목으로 연속해서 소개할 예정이다.

1. DENIAL et al v. OpenAI, Inc.

(1) 캘리포니아 북부 연방지방법원(3:25-cv-05495, 2025.6.30.)

원고는 Catherine Denial, 작가 Lan McDowell, Steven Schwartz 등 3인의 저자들로 구성된 집단소송이며, 피고들은 OpenAI의 모회사 및 계열사와 Microsoft이다.

(2) 사실관계(원고 주장)

- ① 원고 작가들은 기사, 에세이, 기타 어문저작물 등의 저작권자들인데, OpenAI는 원고들의 저작물을 복제하여 LLM 학습에 사용하였다.
- ② OpenAI는 상업적 목적을 위해 수천만 건의 불법(Piracy) 저작물을 그림자도서관(shadow library)으로부터 복제하고, P2P에 의한 토렌트 방식으로 획득, MS는 OpenAI의 핵심 사업 파트너로서 불법 행위에 공동으로 가담하였다.
- ③ OpenAI는 LibGen, Z-Library, Sci-Hub, Internet Archive, Bibliotik과 같은 그림자도서관에 접근하여 저작물을 대량으로 토렌트(torrent) 방식으로 다운로드하고 학습데이터셋으로 편집하였다. 원고들의 저작물이 OpenAI의 데이터셋에 포함되어 있다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 직접 침해: OpenAI와 MS는 LLM과 관련하여, 원고의 저작물 등 방대한 텍스트 자료를 디지털 형식으로 복제(그림자도서관 또는 해적 웹사이트로부터 토렌트 방식으로 수집하는 것도 포함)하였다. AI를 학습, 배포하는 과정에서 저작물을 복제, 저장, 사용한 행위도 복제에 해당한다. 또한 LLM 학습 과정에서도 추가적으로 복제물을 만들었다.
- ② 저작권 간접 침해(대위): MS는 ①OpenAI와의 파트너십 및 OpenAI에 대한 상당한 투자에 근거하여, OpenAI의 침해 행위를 감독하고 통제할 권리와 법적 권리, 그리고 실제 통제할 능력을 가지고 있다, ②가치평가

상승, 라이선스 수익, 기타 상업적 이점 등을 통하여 OpenAI에 의한 침해 행위로부터 직접적인 금전적 이익을 얻었다.

- ③ 접근통제 기술적 보호조치 규정(§1201(a)) 위반: 피고들은 그림자도서관으로부터 원고들의 저작물을 토렌트 방식으로 교환하거나, 로봇배제 프로토콜인 robots.txt 파일을 무시하거나, 기타 접근 관련 보안 조치를 무시하는 등의 방법으로, 원고들의 저작물에 대한 접근을 통제하는 기술적 조치를 무력화하였다.

그림자도서관에 있는 원고들의 저작물에는 접근통제가 적용되었다고 할 수 없다. 그런데 원고들이 접근통제 기술적 조치 규정을 위반했다고 주장하는 것은 자신들 저작물의 정상적인 유통 경로에서는 접근통제가 존재했다거나(예컨대 로그인하거나 이용 대가를 지급하는 등), robots.txt를 무시한 것은 접근통제 무력화에 해당한다고 주장하기 위하여, 또는 공정이용 항변을 원천적으로 봉쇄하거나, MS까지 공동 위반자로 포함시키기 위한 전략(직접 복제를 하지 않은 MS도 접근통제 무력화의 공동 위반자에 해당할 수 있음)이라 볼 수 있다.

- ④ 저작권 관리정보(CMI) 규정(§1202(b)(1)) 위반²⁾: 저자명, 저작물 제목, 출판정보, 이용조건, ISBN 및 저작권 번호 등의 CMI를 제거했다.

2. Kadrey v. Meta Platforms, Inc.³⁾

(1) 캘리포니아 북부 연방지방법원(23-cv-03417-VC, 2023.7.7.)

원고는 Richard Kadrey, Sarah Silverman, and Christopher Golden 등 다수의 작가들로 구성된 집단소송이며, 피고는 Meta Platforms, Inc 이다. 이 사건은 2023.7.7. 제기되었다. 2025.6.25. 법원은 Kadrey 등 13명의 원고에 대하여 AI 학습의 공정이용 여부에 대하여 약식판결을 내렸다.

(2) 사실관계(원고 주장)

- ① Meta는 원고들의 서적이 포함되어 있던 그림자도서관 LibGen과 이를 집계·미러(aggregate, mirror)⁴⁾한 Anna's Archive, 그 유통본을 학습용으로 재패키징한 Books3 데이터셋을 토렌트 방식으로 다운로드하여 AI 모델인 Llama의 학습데이터로 사용하였다.

1) 로봇배제표준 또는 'robots.txt 파일'은 웹사이트 운영자가 검색엔진 등의 크롤러(crawler)나 웹로봇에게 웹상의 콘텐츠에 접근 허용 또는 금지 여부 및 범위를 알리는 자발적 준수 표준(규약) 또는 파일임. '자발적'인 것이므로 이를 무시하여 크롤링하는 것은 가능하고, 암호화나 인증 없이 접근할 수 있으므로, 이러한 무시가 기술적 보호조치 무력화에 해당하기는 어려울 것임. 다만 이 사건에서 원고는 robots.txt 뿐만 아니라 DRM 비밀번호(주장은 하고 있으나 구체적으로 밝히지는 않고 있음), 스크래핑을 금지하는 이용약관, 접근 관련 보안조치 등과 함께 기술적 보호조치 무력화를 주장하고 있음.

2) 미국의 AI 저작권 케이스에서는 CMI가 DMCA에 의하여 보호된다거나, DMCA가 CMI 제거를 금지한다거나, 피고가 DMCA 규정을 위반하였다는 것 등이 언급되고 있음. DMCA(Digital Millennium Copyright Act)는 1998년 미국 저작권법에 인터넷을 반영하기 위한 개정 입법으로서, OSP 면책(§512), 기술적 보호조치(§1201), 저작권 관리정보(§1202) 등의 규정을 신설한 입법(Public Law No: 105-304)임. 예컨대 DMCA 제103조 (a)항은 "미국 저작권법은 끝 부분에 다음의 새로운 장(chapter)을 추가함으로써 개정된다"고 하면서, Chapter 12에 해당하는 제1201 내지 제1205조(현재의 기술적 보호조치 및 권리관리정보에 관한 규정)를 나열하고 있음. 곧 1998년 미국 저작권법은 DMCA에 의하여 현재의 Chapter 12가 추가된 것임. 따라서 '미국 저작권법 제1202조'가 정확하고, 'DMCA 제1202조'는 잘못된 표현임.

3) 이 케이스에 대하여 좀 더 상세한 것은 이대희, "[이슈리포트] 2025-3-[미국] 앤스로픽 및 메타 케이스의 비교 분석" 참조.

4) 집계는 여러 출처의 콘텐츠를 하나의 접근, 검색 구조로 통합하는 것이고, 미러는 특정 출처의 데이터 전체를 그대로 복제하여 저장하는 것을 의미.

② Meta는 저작권이 있는 자료 등 특정 텍스트를 암기(memorize)하거나 출력하지 못하도록 모델을 사후 학습시켰고, 이런 조치는 성공적이었다. 곧 Meta의 모델은 원고들의 서적에서 50개를 초과하는 단어 및 문장부호(토큰)를 생성하지 않았다.

(3) AI 학습의 공정이용 여부 (2025.6.25. 판결)⁵⁾

- ① 공정이용의 첫째 판단 요소: Meta의 저작물의 이용은, 다양한 텍스트를 생성하고 광범위한 기능을 수행하는데 사용될 수 있는 혁신적 도구인 LLM을 학습시키기 위한 것으로서, 높은 수준의(highly) 변형적 이용(transformative use)에 해당한다.⁶⁾ 다운로드를 궁극적으로 Llama를 학습시키는 것이 매우 변형적이라는 목적에서 고려되어야 하고, 서적을 이용한 궁극적인 목적이 변형적이므로 서적의 다운로드도 변형적이다.
- ② 둘째 요소: 원고의 서적들은 고도의 표현적인 저작물로서 둘째 요소는 원고에게 유리하게 작용하지만, 전체 공정이용 여부 분석에서 큰 의미를 가지지 않는다.
- ③ 셋째 요소: Meta가 원고들의 서적을 전부 복제하였으나, 복제한 분량은 변형적 목적과의 관계에서 볼 때 합리적인 것으로서, Meta에게 유리하게 작용한다.
- ④ 넷째 요소: ①Llama는 적대적 프롬프트를 입력하더라도 원고들 서적의 50단어를 초과하여 생성하지 않는 등, Llama 생성 결과물은 원고들 서적의 잠재적 시장이나 가치에 의미가 있을 정도의 효과를 미칠 위험이 되지 않는다; ②‘학습데이터 이용허락 시장’은 이론적 시장(theoretical market)으로서 잠재적 시장이 되지 않는다(법원은 저작권자가 사용료를 받을 수 있었다고 하여 시장에 대한 영향이 있었다고 인정하면, 공정이용 판단은 항상 저작권자에게 유리하게 작용하는 순환논리에 빠진다고 지적)⁷⁾; ③저작권을 직접 침해하지는 않지만, 침해에 해당되지 않는 결과물이 원저작물과 경쟁함으로써 원저작물의 시장이 희석(dilution)될 가능성이 있으나 원고가 이를 입증하지 못함으로써, 시장의 희석을 인정할 수 있다.

(4) 공정이용 여부 결론

Meta는 원고들의 서적을 학습데이터로 이용한 것이 저작권 침해라는 주장에 대하여 공정이용 항변을 하였고, 법원은 Meta가 학습데이터 이용이 공정이용이라는 약식판결⁸⁾을 인정받을 수 있다고 판단하였다.

5) Kadrey v. Meta Platforms, Inc., 23-cv-03417-VC(N.D. Cal. 2025.6.25).

6) 변형적 이용 여부는 공정이용을 판단하기 위한 첫째 요소를 결정하는 핵심 요소인데, 대체로 원저작물과 다른 방식으로 이용하거나 다른 목적(different purpose)을 위하여 이용하는 것을 의미함. 원저작물을 단순히 재포장하거나 재공표하는 식으로 이용하는 것은 단순히 원저작물의 “대상을 대체(supersede)하는 것은 변형적 이용이 되지 않으나, 원저작물에 가치를 추가하는 경우는 변형적 이용이 될 수 있음.

7) ‘순환논리’란 논리적 논증에서 특정한 주장을 하면서 그 주장의 근거로 다시 그 주장을 사용하는 오류를 의미함. 예컨대 학습데이터 이용이 공정이용이 아니라는 것을 입증하여야 하는데, 이를 위해서는 해당 저작물 시장이 실제 또는 잠재적 시장이 아니라는 것을 입증하여야 함. 학습데이터 이용허락 시장을 잠재적 시장으로 인정하는 경우 항상 공정이용이 부인되는 결과가 됨. 곧 공정이용이 아니므로 저작권 침해가 성립한다고 입증하여야 하는데, 잠재적 시장 부인에 따른 침해 부인으로 입증하는 식이 됨으로써, 순환논리에 빠지게 됨.

8) 약식판결(summary judgment)은 ‘주요 사실관계(material fact)’에 관하여 ‘진정한 다툼’이 없고(no genuine dispute) 따라서 법률적 판단만으로 결론을 내릴 수 있는 경우, 당사자의 청구에 따라 내리는 판결임(FRPC §56). 약식판결 청구의 상대방은 주요 사실관계에 관하여 진정한 다툼이 있다는 것을 입증하여야 약식판결 원용을 방어할 수 있음. 이 사건에서 원고는 Meta의 행위가 공정이용이 아니라는 원고 승소의 약식판결을 청구하고, Meta는 공정이용이라고 판단해 달라는 피고 승소의 약식판결을 청구하였음. 교차 약식판결 청구의 경우, 각 당사자는 자신의 주장은 다툼이 없는데, 상대방의 주장은 다툼이 있다는 것을 입증하여야 하고, 법원은 각 청구를 개별적으로 판단하여 증거와 법리상 어느 쪽이 더 타당한지 우열을 가리게 됨. 원고가, Meta에 대한 약식판결 인용을 배척하려면, 공정이용의 네 번째 요소를

이 사건은 집단소송으로 제기되었으나 집단으로 인증받기에 적절하지 않았고 이에 대한 판결은 13명의 저작권자들에게 한정된다. 곧 처음에는 Richard Kadrey 3명의 저작권자들이 집단소송을 제기하였는데 이후 소장을 수정하여 총 13명으로 증가하였고, 2025.6.25. 판결은 이들 13명에 대한 것이다.

원고(서적의 저작권자)가 Meta의 행위가 저작권 침해에 해당하고 공정이용이 아니라는 원고 승소의 약식판결을 청구하고, Meta(피고)가 공정이용이라는 피고 승소의 약식판결을 청구하였다. 공정이용의 첫 요소와 관련하여 AI 학습을 위한 학습데이터 이용이 고도로 변형적이었으므로, 원고들은 공정이용의 네 번째 판단 요소에서 결정적으로 유리한 입증이 필요하였는데, 입증을 하지 못함으로써 공정이용이라는 Meta의 약식판결 청구가 인용된 것이다. 만약 원고가 시장 희석을 입증하였다 가정한다면, 첫째 요소는 Meta에게 유리하게 작용하고, 넷째 요소는 원고에게 유리하게 작용하는데, 이에 대하여 법원이 어떻게 판단할지는 미지수이다.

(5) 저작권 관리정보 규정 위반 여부(2025.6.27. 판결)⁹⁾

- ① 원고는 CMI 제거 금지에 관한 규정(§1201(b)(1)) 위반을 주장하였고, Meta는 위반이 아니라는 약식판결을 청구하였다.
- ② 법원은, Meta의 복제 행위가 침해가 아니고 따라서 CMI를 제거한다고 하여 저작권 침해행위를 용이하게 하는 것이 될 수 없으므로, CMI 제거에 관한 규정을 위반하지 않는다는 것에 대한 Meta의 약식판결 청구를 인용하였다.

3. Bird v. Microsoft Corporation

(1) 뉴욕 남부 연방지방법원(1:25-cv-05282, 2025.6.25.)

원고는 풀리처상 수상작가 Kai Bird, Jia Tolentino, Daniel Okrent 등 저자들로 구성된 집단소송이며, 피고는 Microsoft Corporation이다.

(2) 사실관계(원고 주장)

- ① MS는 'The Pile' 데이터셋의 하위 데이터셋을 이용하여 Megatron LLM¹⁰⁾을 학습시켰고, The Pile에는 데이터셋인 Books3(약 20만 권의 서적)가 포함되어 있다.
- ② MS는 학습시키는 과정에서 데이터셋에 포함된 저작자들에게 보상하지 않았고, 도서 판매 및 라이선스 수익을 박탈하였다.

판단함에 있어서 저작물 이용으로 인하여 시장에 영향이 있는지 여부에 대하여 진정한 다름이 있다고 입증하여야 하고, 이러한 입증은 저작물의 판매 감소, 이용허락 시장 침해, 소비자 대체 등에 대한 경험적 증거(empirical evidence)에 의하여야 하는데, 원고는 이러한 입증을 하지 못한 것임.

9) Kadrey v. Meta Platforms, Inc., 23-cv-03417-VC (2025.6.27.).

10) MS는 Turing 계열(MS가 Research가 개발한 대규모 AI 모델들의 제품군)의 개발자인데, 처음에는 'Turing-NLG'이라는 Turing LLM 데모 버전을 공개하였다가, 2021.10. 이보다 훨씬 큰 모델인 'Megatron-Turing Natural Language Generation model'을 공개하였음. Megatron 또는 Megatron LLM은 바로 이 모델을 의미함.

(3) 침해(위반) 주장 및 청구 사항

MS는 ①해적 사이트에서 원고의 서적을 다운로드받고 ②복제된 저작물로 학습을 시킴으로써, 복제권을 침해하였다.

4. Bartz v. Anthropic PBC¹¹⁾

(1) 캘리포니아 북부 연방지방법원(3:24-cv-05417, 2023.7.7.)

원고는 Andrea Bartz, Charles Graeber, Kirk Wallace Johnson 등으로 구성된 집단소송이며, 피고는 Anthropic이다. 이 사건은 2023.7.7. 제기되었고(3:24-cv-05417), 2025.6.23. AI 모델 학습, 인쇄본 서적의 디지털화, 학습데이터 다운로드에 대한 공정이용 판결이 이루어졌고, 2025.12.15. 현재 화해 절차가 진행 중이다.

(2) 사실관계

- ① 피고 Anthropic PBC는 ①그림자도서관인 Library Genesis(LibGen)로부터 최소 500만 권, Pirate Library Mirror(PiLiMi)로부터 최소 200만 권의 도서를 다운로드받았다. 또한 ②수백만 권의 인쇄본 도서(종교)를 구매한 뒤, 스캔하여 디지털 파일로 제작(종이 원본은 폐기). 이들 파일들로 구성된 라이브러리(central library)의 데이터를 이용하여 AI 모델을 학습시켰다.
- ② Anthropic은 자신의 저작물 이용이 공정이용이라는 약식판결을 청구하였다.

(3) 공정이용 여부에 대한 법원의 판결(2025.6.23.)

- ① 해적사이트로부터의 다운로드: 공정이용에 절대로 해당되지 않는다.
- ② 인쇄본 서적의 포맷 변경(디지털화): 공정이용에 해당한다.
- ③ AI 모델 학습: (a)학습은 변형적 이용에 해당하여 학습이 공정이용이 되는 방향으로 작용하는 중요한 역할을 한다. (b)공정이용 두 번째 요소는 공정이용이 되지 않는 방향으로 작용한다. (c)셋째 요소는 공정이용이 되는 방향으로 작용한다. (d)④학습데이터와 동일한 결과물을 AI가 생성하지 않으므로 원고 서적의 잠재적 시장이나 가치에 의미가 있거나 상당한 영향을 미칠 위협이 없다. ④학습데이터 이용허락 시장은 이론적 시장에 해당하여 잠재적 시장이 될 수 없다. ⑤LLM에 의하여 저작물과 경쟁하는 결과물이 생성된다는 주장은 학생들의 작문 학습에 의하여 원저작물과 경쟁하는 저작물이 만들어진다는 것과 동일한 것으로서, 이러한 경쟁은 저작권법이 보호하고자 하는 것이 아니며, 학습은 오히려 창작을 증진한다. 따라서 네 번째 요소도 공정이용이 되는 방향으로 작용, 두 번째 요소를 제외하는 3개 요소 모두 공정이용이 되는 방향으로 작용하는데, 법원은 학습을 위한 원고 저작물 이용이 공정이용이라고 판시하였다.

(4) 화해

- ① 분쟁 당사자들은 2025.8.26. 화해에 의하여 사건을 종결시키기로 하였고 2025.8.26. 화해안을 법원에 제출하였다. 집단소송에서의 화해는 법원의 승인을 얻어 확정(연방민사소송규칙(FRPC) §23(e)). 2025.12.15. 현재

11) 이 케이스에 대하여 좀 더 상세한 것은 이대희, 앤스로픽 및 메타 케이스의 비교 분석(저작권동향, 2025-3) 참조.

화해안을 승인하기 위한 절차가 진행 중이다.

- ② 화해안은 Anthropic이 15억 달러의 화해 기금을 마련하여 저작권자에게 지급하고, 화해 기금에서 배상받을 저작권자 집단¹²⁾은 Anthropic이 LibGen과 PiLiMi에서 다운로드받은 서적의 저작권자들(서적은 약 50만여 권 예상)이고, (a)LibGen과 PiLiMi에서 다운로드받은 저작물 파일과 (b)토렌트로 다운로드받은 복제물에서 파생하는 모든 복제물을 파기하는 것 등을 내용으로 한다.

Anthropic의 공정이용 여부와 관련하여 해적판 다운로드를 공정이용이 아닌 것으로 판시되었으므로, 소송이 계속 진행되어 저작권 침해가 확정될 경우, 앤스로픽은 50만여 권의 서적에 대한 법정손해배상을 부담할 가능성이 있다(고의에 의한 경우, 저작물 당 15만 달러까지 될 수 있음). 또한 저작권자들도 서적당 3,000 달러를 배상 받게 되는데, 이는 비교적 큰 금액에 해당하는 것이어서, 양자 간의 화해에 의한 종결로 합의한 것으로 보인다

5. Raw Story Media, Inc. v. OpenAI Inc.¹³⁾

(1) 뉴욕 남부 연방지방법원(1:24-cv-01514, 2024.2.28.)

원고는 온라인 뉴스 웹사이트인 'Raw Story'를 운영하는 기업 Raw Story Media, Inc.와 뉴스 및 비평 사이트인 'AlterNet'을 운영하는 AlterNet Media, Inc.이다. 피고들은 OpenAI의 모회사 및 계열사이다.

(2) 사실관계(원고 주장)

- ① 언론사인 원고 Raw Story의 저널리즘 저작물은 저자, 저작물 명칭, 저작권 정보(저작권 관리정보, CMI)와 함께 공중에게 제공된다. OpenAI의 ChatGPT는 저작물의 상당한 부분을 모방하면서도, 해당 저작물에 포함된 CMI를 전혀 제공하지 않는 응답을 이용자들에게 제공하였다. 원고의 수천 개 저작물은 CMI가 제거된 상태로 ChatGPT를 학습시키는데 사용된 학습 데이터셋에 포함되었다.
- ② OpenAI는 CMI가 제거된 상태로 원고들의 저작물을 학습 데이터셋에 포함시킴으로써, ChatGPT 및 ChatGPT 이용자에 의한 저작권 침해를 가능하게 하거나, 침해를 용이하게 하거나, 은폐할 것임을 알거나 알 만한 합리적인 이유가 있었고, 이에 따라 CMI 제거를 금지하는 규정(§1201(b)(1))을 위반하였다.

(3) 법원 판결 내용

2024년 11월, 뉴욕 남부 연방지방법원은 원고가 구체적인 피해를 입증하지 못했다는 이유(당사자 적격 결여)로 소송을 기각했다. 원고들은 이 결정에 불복하여 항소했으며, 현재 이 사건은 제2 연방항소법원에서 계류 중이다.

6. Justice v. Uncharted Labs, Inc.

12) 약식판결을 내린 이후 2025년 7월 17일 법원은 앤스로픽이 '그림자도서관'인 Library Genesis('LibGen')와 Pirate Library Mirror('PiLiMi')에서 다운로드받은 서적의 저작권자들을 '집단(class)'으로 인정하는 판결을 하였음. Bartz v. Anthropic PBC, Order on Class Certification (C 24-05417 WHA. N.D.Cal. 2025.7.17.).

13) 이 케이스에 대하여 좀 더 상세한 것은 차상욱, 언론사 Raw Story와 AlterNet이 OpenAI를 상대로 제기한 저작권 침해 소송 각하(저작권 동향, 2024-32) 참조.

(1) 뉴욕 남부 연방지방법원(1:25-cv-05026, 2025.1.18.)

원고는 Anthony Justice, 5th Wheel Records 등으로 구성된 집단소송이며, 피고는 UNCHARTED LABS, INC., d/b/a Udio.com 등이다.

(2) 사실관계(원고 주장)

- ① 원고는 컨트리 음악 가수 및 음반제작사이고, 피고 Uncharted Labs(Udio라는 명칭으로 행위)는 AI 음악 생성기(generator)를 학습시키기 위하여 원고들의 음향저작물(sound recording)을 이용하였다.
- ② Udio의 다운로드(데이터 수집) 및 학습 과정은 저작물 복제에 해당한다.
- ③ Udio의 복제는 공정이용에 해당하지 않는다. 곧 Udio는 상업적 이익을 위하여 저작물을 이용하고, 원고와 완전히 경쟁 관계에 있는 결과물을 생성한다. 원고 저작물은 창의적 표현이고, 광범위하게 원고 저작물의 복제가 이루어지고, 원고 저작물의 가장 중요한 부분의 복제가 이루어졌다. Udio의 저작물 이용은 원고 음악의 시장과 가치에 중대한 위협을 가하고 있고, 저작물과 실질적으로 유사한 결과물을 생성하도록 모델을 학습시키는 행위는 변형적이지 않다.

(3) 침해(위반) 주장 및 청구 사항

- ① Udio는 원고들의 음악을 이용하여 생성형 AI 모델을 학습시키는 행위는 복제권 침해에 해당한다.
- ② Udio는 원고들의 음악을 각색함으로써 2차적저작물 작성권을 침해하였다.

7. Justice et al v. Suno, Inc.**(1) 매사추세츠 연방 지방법원(1:2025cv11739, 2025.6.14.)**

제소한 법원과 피고(Suno)가 다르다는 점을 제외하고 Justice v. Uncharted Labs, Inc. 사건(1:25-cv-05026, 2025.1.18.)과 동일하다.

8. Ziff Davis, Inc. et al v. OpenAI, Inc.**(1) 뉴욕 남부 연방지방법원(1:2025cv04315, 2025.4.24.)**

원고는 Ziff Davis, Inc. 및 산하 브랜드(PCMag, Mashable, IGN, Everyday Health, Lifehacker 등) 등으로 구성된 집단소송이며, 피고는 OpenAI, Inc. 및 관련 법인이다.

(2) 사실관계(원고 주장)

- ① 원고는 Ziff Davis(언론사), IGN Entertainment, Everyday Health Media이다. OpenAI는 LLM 소프트웨어 및 LLM 기반 제품과 서비스를 구축하고 운영하기 위하여, Ziff Davis 저작물을 의도적으로 복제하고 파생물(derivative)¹⁴⁾을 생성하였다.

14) 미국 저작권법상 2차적저작물(derivative work)은, 번역, 음악적 편곡, 드라마화, 허구화, 영화화, 음향저작물, 미술 복제, 요약, 축약, 그 밖

- ② OpenAI는 AI 모델을 개발하고 서비스를 제공하기 위하여 사전학습 이외에, 후속 학습(post-training), 미세조정 등 다양한 방법을 사용. 2018년 이후 GPT-1, GPT-2, GPT-3, ChatGPT/GPT-3.5, GPT-4, GPT-4o, GPT-4o-mini, o1 mini, o1, o1-pro, GPT-4.1, GPT-4.5, o3, o3-mini, o4-mini 등 일련의 LLM을 출시하였다.
- ③ OpenAI의 LLM 학습은 창작적 표현을 압축시키고 결과물 생성은 압축을 해제하는 것으로서, LLM은 그 자체가 복제물을 포함하는 것이다. OpenAI는 RAG이라는 보조기법을 사용하여 웹과 같은 외부 정보 저장소를 참조하고 검증하여 결과물 제공한다.
- ④ OpenAI는 (a)원고의 웹사이트에서 콘텐츠를 스크래핑하여 학습, 데이터셋에 복제하고, 이 과정에서 OpenAI는 기술적 보호조치를 무력화하고, 저작권 관리정보(CMI)를 제거하고, (b)학습 과정에서 원고 저작물의 복제물이나 파생물을 인코딩(encoding)하여 LLM 내부에 생성하고, (c)LLM 결과물에서 원고 저작물의 복제물 또는 파생물을 생성하고, (d)원고 저작물의 복제물을 포함하는 LLM 생성물을 저장하여 학습 및 제품 개선에 사용하고, (e)RAG 기능을 활용하는 과정에서 추가적인 LLM 입력 및 결과물을 생성하기 위하여 원고의 사이트에서 콘텐츠를 스크래핑하고 복제하고 있다(원고는 robots.txt. 적용).
- ⑤ OpenAI는 데이터셋을 생성할 때 CMI를 제거한 뒤 텍스트 콘텐츠만을 추출하는데, CMI 제거는 침해를 은폐하고, 자신 및 타인이 침해를 더 쉽게 하도록 하기 위한 것이다. 또한 API¹⁵⁾를 통하여 LLM에 대한 접근을 제공하면서 CMI가 제거된 원고 저작물의 복제물을 고객들에게 직접 배포하고 있다.

(3) 침해(위반) 주장 및 청구 사항

- ① 복제에 의한 저작권 침해: OpenAI는 원고 웹사이트에서 콘텐츠를 스크래핑하고 제3자 데이터셋으로부터 원고의 콘텐츠를 복제함으로써, 원고의 수백만 건의 기사 및 기타 콘텐츠를 학습 데이터셋에 복제하였다. LLM 학습을 위하여 원고 저작물의 수백만 건 복제물을 포함하는 학습 데이터셋을 저장·처리·복제하였다(LLM 자체에 복제하는 것 포함).
- ② 결과물에 의한 복제에 따른 침해: 원고 저작물의 복제물 및 파생물을 포함하는 결과물을 배포하기 위하여 원고 저작물을 복사하고 복제,¹⁶⁾ 원고 저작물을 시스템, 소프트웨어, 제품 및 서비스에 통합(incorporate)하였다.
- ③ 간접 침해(기여 침해): 사용자들의 직접 침해를 인식하고 이러한 침해에 실질적으로 기여하였다.
- ④ 기술적 보호조치 무력화: 원고들은 저작물에 대한 접근을 효과적으로 통제하기 위하여 robots.txt 등 접근제한 조

에 저작물이 재구성되거나 변형되거나 각색될 수 있는 모든 형태와 같이, 하나 이상의 기존 저작물을 기초로 한 저작물로 정의됨(\$101). AI 저작권 소송에서 원고는 2차적저작물보다는 파생물(derivative)이라는 용어를 많이 사용하고 있음, AI 모델은 학습데이터(원저작물)와 동일한 것과 유사한 것을 생성하므로, 이들 양자를 모두 지칭하기 위하여 파생물이라는 용어를 사용하는 것으로 보임.

15) API(Application Programming Interface)는 AI 모델이나 AI 기능을 다른 소프트웨어가 정해진 방식으로 호출·사용할 수 있도록 만든 공식적인 접점(인터페이스)을 의미함. 제3자는 모델을 직접 보유하거나 학습시키지 않고 API 호출을 통하여 기능만 빌려쓰게 됨.

16) 미국 저작권법상 복제는 'reproduce(reproduction)'인데, 원고는 복제와 함께 복사(copy)라는 용어까지 많이 사용하고 있음. 복사는 사실적 행위를 넓게 지칭하는 용어이고 복제는 법률적인 엄격한 용어이므로, 원고는 가능한 한 피고의 불법행위를 광범위하게 표시하기 위하여 복사라는 용어까지 사용하는 것으로 보임. 예컨대 AI 모델의 파라미터(parameter)가 복제물에 해당하는가 문제될 수 있는데, 원고는 이를 복사라고 표현함으로써 법률적인 복제에 해당하는 것처럼 서술하기 위한 전략이라 할 수 있음.

치를 취하였는데, OpenAI는 이를 무력화하여 저작물에 접근하였고, 이에 따라 접근통제 기술적 보호조치 무력화 금지규정(§1201(a)(1))을 위반하였다.

- ⑤ 저작권 관리정보 제거: 원고는 저작물에 CMI(저작권 표시, 바이라인(bylines),¹⁷⁾ 저작자 정보, 저작자 명칭, 저작물 명칭, 이용 조건 등)를 표시하였는데, OpenAI는 데이터셋을 구축하기 위하여 원고의 CMI를 고의로 제거. OpenAI는 자신이나 사용자에게 의한 잠재적 침해를 가능하게 하거나, 용이하게 하거나, 은폐할 것이라는 인식과 의도를 가지고 CMI를 제거하거나 변경하였으며, 이에 의하여 원고 저작권을 모호하게 하고 부당한 이용을 조장하였다.
- ⑥ 저작권 관리정보가 제거된 저작물의 배포: OpenAI는 CMI가 제거된 원고 저작물과 파생물을 복제·생성하고 이를 제3자에게 배포하였다. 이는 CMI가 제거·변경되었다는 것을 인지하고서 저작물을 배포·수입·공연하는 것을 금지하는 규정(§1201(b)(3), 한국 저작권법 §104조의3① iii에 해당)을 위반하는 것에 해당한다.

9. Millette v. OpenAI, Inc.

(1) 캘리포니아 북부 연방지방법원(3:24-cv-04710-TSH, 2024.8.2.)

원고는 DAVID MILLETTE가 전체 집단의 대표 원고로 소송을 진행한 집단소송이며, 피고는 OpenAI, Inc. 및 관련 법인이다.

(2) 사실관계(원고 주장)

- ① 원고들은 유튜브 사용자이자 영상 제작자들이다.
- ② OpenAI의 학습데이터를 구성하는 많은 부분은 원고가 생성하고 업로드한 영상물이고, OpenAI는 전사(transcribe, 음성·영상에 포함된 말을 그대로 텍스트로 옮기는 행위)하여 모델 학습과 미세조정에 사용하였다.

(3) 침해(위반) 주장 및 청구 사항

주법에 기반하여 부당이득, 원상회복, 불공정 경쟁을 주장하고 있다.

원고들이 저작권 침해를 주장하지 않은 것은 '전사'의 법적 지위가 다소 모호하므로 OpenAI에 의한 공정이용의 항변을 회피하기 위한 것으로 보인다. 전사는 음성을 문자로, 영상 속의 음성을 자막으로, 오디오 파일을 텍스트 문서로 바꾸는 것처럼, 기존 콘텐츠를 다른 형식으로 옮겨 적는 것을 의미한다. 전사는 저작물의 복제나 2차적저작물 작성에 해당할 수 있다. 그러나 AI 모델 이용자가 전사를 지시할 수도 있고, 전사할 권리를 보유할 가능성도 있으며, 전사가 공정이용에 해당하거나 저작권자가 전사에 대하여 묵시적으로 허락하였다고도 볼 수 있다. 원고가 저작권 침해가 아니라 주법 위반을 주장한 것은 이러한 이유에 기인하는 것으로 보인다.

10. Tremblay v. OpenAI, Inc.

17) 'Written by John Smith'와 같이 신문·잡지·온라인 기사 등에서 해당 저작물을 누가 작성했는지를 나타내는 표기.

(1) 캘리포니아 북부 연방지방법원(3:23-cv-03223-AMO, 2023.6.28.)

원고는 Paul Tremblay가 전체 집단의 대표 원고로 소송을 진행한 집단소송이며, 피고는 OpenAI, Inc. 및 관련 법인이다.

(2) 사실관계(원고 주장)

- ① OpenAI의 학습 데이터셋 BookCorpus는 미출판 소설들을 호스팅하는 Smashwords.com에서 서적들을 복제하였고, GPT-3 학습 데이터셋의 15%는 Books1 및 Books2인데, 각각 63,000권 및 294,000권의 서적이 포함되어 있을 것으로 추정된다.
- ② ChatGPT가 원고 서적을 정확하게 요약하는 결과물을 제공하는 것은 이들 서적들이 학습데이터의 일부로서 LLM에 입력하여 처리되었기 때문이다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 직접 침해: OpenAI는 원고들의 허락 없이 언어 모델의 학습 과정에서 원고들의 서적을 복제하고, AI 결과물로서 2차적저작물 생성하였다.
- ② 저작권 간접(대위) 침해: LLM 모델의 결과물 생성을 통제할 권리와 능력을 가지고 있고, 모든 결과물은 대위 침해 행위를 구성한다.
- ③ 저작권 관리정보 제거: 원고 저작물에는 저작권 표시, 저작물 제목, 저작권자의 명칭 및 기타 식별 정보, 이용 조건 등의 CMI가 표시되어 있는데, ①OpenAI는 CMI를 제거하고, ②2차적저작물에 해당하는 결과물을 CMI 없이 배포함으로써 CMI에 관한 규정(§§1202(b)((1), (3))을 위반하였다.

11. OpenAI 관련 소송들의 단일 법원 집중

(1) 소송의 관할 집중

미국의 연방 '다수 관할 소송 재판부(Judicial Panel on Multidistrict Litigation, 이하 'JPML')'는 2025.4.3. OpenAI 및 MS를 피고로 하여 캘리포니아 및 뉴욕의 연방지방법원에 제기된 사건들이 공통된 사실관계를 가지고 있고, 소송이 이루어지기 전의 절차를 조정·통합하기 위하여 해당 사건들을 '뉴욕 남부 연방지방법원(SDNY)'으로 이송하는 명령(transfer order)을 하였다.¹⁸⁾ JPML은 미국 연방법원 관할구역에 여러 민사소송이 계류 중인 경우, 재판 절차를 조정·통합하기 위한 특별 재판부(panel)로서, 이러한 소송들이 공통의 쟁점을 가지고 있으므로 단일 연방지방법원으로 이송할 것인지를 결정한다.¹⁹⁾ JPML의 SDNY 이송 명령은 OpenAI가 캘리포니아 북부지방법원으로 집중할 것을 청구한 것에서 시작되었다. JPML은 첫째, OpenAI와 Microsoft가 저작권자의 동의나 보상 없이 저

18) JPML, In Re: OpenAI, Inc. Copyright Infringement Litigation, MDL No. 3143(2025.4.3.).

19) 연방 지방법원에 관하여 규정하는 법규(28 USC Ch. 87 Part IV, 관할권)에 규정되어 있음. 이 재판부는 연방대법원장에 의하여 임명되는 7명의 연방 판사(13개의 연방 고등법원 관할구역 중에서 각기 다른 구역이어야 함)로 구성됨. 병합 결정은 공통의 사실관계 쟁점의 존재, 당사자와 증인의 편의, 공정하고 효율적 소송 진행 등의 요건을 충족하여야 함. 28 USC §1407.

작물을 이용하여 GPT-4와 같은 LLM을 학습시켰고, 이 LLM들이 OpenAI의 ChatGPT 및 Microsoft의 Bing Chat(현재 Copilot으로 명칭 변경)과 같은 생성형 AI 제품의 기반이 된다는 주장에서 사실적 쟁점이 동일하고, 둘째, 각각의 사건들은 피고들이 자신들의 LLM을 어떻게 학습시키고 설계했는지에 관하여 중첩되고, 복잡하며, 방대한 증거개시 절차(discovery)를 수반하게 될 것이고, 이 사건들 전반에 걸쳐 중복되는 전문가들이 등장할 가능성도 높다고 판단하였다. 따라서 JPML은 관할 집중에 의하여 중복된 증거개시 절차를 제거하고, 특히 집단(class) 인증²⁰⁾에 대한 불일치를 방지하며, 당사자, 소송대리인, 사법부의 자원을 절약 등의 편의를 위한 것이라는 명령의 근거를 제시하였다.

(2) 관할 집중되는 사건(총 12개)

1) 캘리포니아 북부 연방지방법원

- ① Tremblay et al. v. OpenAI, Inc. et al. (3:23-cv-03223, 2023.6.28.)
- ② Silverman et al. v. OpenAI, Inc. et al. (3:23-cv-03416, 2023.7.7.)
- ③ Chabon et al. v. OpenAI, Inc. et al. (3:23-cv-04625, 2023.9.8.)
- ④ Millette v. OpenAI, Inc. et al. (5:24-cv-04710, 2024.8.2.)

2) 뉴욕 남부 연방지방법원

- ① Authors Guild et al. v. OpenAI, Inc. et al. (1:23-cv-08292, 2023.9.19.)
- ② Alter v. OpenAI, Inc. (1:23-cv-10211, 2023.11.21.)
- ③ The New York Times Company v. Microsoft Corporation et al. (1:23-cv-11195, 2023.12.27.)
- ④ Basbanes v. Microsoft Corporation (1:24-cv-00084, 2024.1.5.)
- ⑤ Raw Story Media, Inc. et al. v. OpenAI, Inc. et al. (1:24-cv-01514, 2024.2.28.)
- ⑥ The Intercept Media, Inc. v. OpenAI, Inc. (1:24-cv-01515, 2024.2.28.)
- ⑦ Daily News LP v. Microsoft Corporation (1:24-cv-03285, 2024.4.30.)
- ⑧ The Center for Investigative Reporting, Inc. v. OpenAI, Inc. (1:24-cv-04872, 2024.6.27.)

(3) 관할 집중에 따른 절차들

12개 사건이 하나의 사건으로 병합된다는 것은 아니고, 단일의 판사(Stein 판사)가 증거 발견, 증언 녹취 조율, 전문가 관리, 기각 청구, 집단 인증, 공정이용 법리 등을 통합 관리하게 된다. 12개 사건들은 뉴욕 남부 연방지방법원(SDNY)에서 여전히 단일 사건으로 진행되고 있다.

12. The New York Times Company v. Microsoft Corporation

²⁰⁾ AI 저작권 소송은 집단소송의 형태로 많이 제기되어 있는데, 집단소송에서는 어떠한 원고가 집단에 포함되는지 별도의 판단을 필요로 함. 예컨대 Bartz v. Anthropic PBC 케이스에서는 AI 학습 및 해적 사이트 다운로드의 공정이용 여부에 대한 판결에 이어, 저작권자를 인증하는 판결을 별도로 하였음. 이 케이스에 대하여 좀 더 상세한 것은 이대희, 앤쓰로픽 및 메타 케이스의 비교 분석(저작권동향, 2025-3) 참조.

(1) 뉴욕 남부 연방지방법원(1:23-cv-11195, 2023.12.27.)

원고는 미국의 저명한 일간지 발행사인 The New York Times이고, 피고는 오픈AI의 최대 투자자이자 파트너사인 Microsoft 사로 오픈AI 저작권 침해 소송이다.

(2) 사실관계(원고 주장)

- ① The New York Times(이하 NYT)는 피고들이 원고들의 콘텐츠를 허락 없이 사용하여 모델과 도구를 개발하고 있다는 사실을 인지한 후, 이에 대해 이의를 제기하고 콘텐츠 사용에 대하여 협상하려 하였으나 협상에 따른 해결에 이르지 못하는 못하였다.
- ② 피고 MS는 OpenAI와의 파트너십(partnership)에 의하여 NYT의 저작물을 복제하고 OpenAI의 생성형 AI 모델을 훈련하는 데 사용된 클라우드컴퓨팅 서비스를 제공하였고, 양자는 모델들의 개발 과정에서 실질적인 기술협력을 하였다. MS는 NYT 저작물의 복제물을 포함하고 이를 처리하여 입력시킨 최신 생성형 AI 모델의 사본을 보유하거나, 이에 대해 우선적 접근권을 취득하고 있는데, 이러한 모델들을 사용하여 자사의 제품 이용자들에게 침해 콘텐츠 및 허위 정보를 제공하고 있다.
- ③ 피고들은 트랜스포머(transformer) 모델을 매우 방대한 양의 데이터로 사전 학습시키고, 특정 지도(supervised) 데이터셋으로 모델을 미세조정 하였다. OpenAI는 GPT-4의 학습데이터를 공개하지 않았지만, 이전 버전에 사용된 Common Crawl, WebText, WebText2 등의 데이터셋을 고려하면, 수백만 건의 NYT 저작물을 사용했을 가능성이 높다.
- ④ NYT의 복제물 및 파생물은 OpenAI의 GPT LLM 자체에 파라미터 및 인코딩된 형태로 암기되어 있다.
- ⑤ OpenAI의 GPT가 생성하는 결과물은 NYT 저작물을 전시하는 것에 해당한다.
- ⑥ GPT LLM을 기반으로 구축된 합성 검색 애플리케이션(Bing Chat²¹) 및 Browse with Bing for ChatGPT²²) 등은 (학습 데이터셋에는 포함되지 않은) NYT의 콘텐츠, 광범위한 발췌물 및 패러프레이즈한 것(paraphrase, 원문의 표현은 바꾸지만 원문이 전달하는 내용을 실질적으로 재현하는 것)을 전시한다.
- ⑦ Bing Chat과 ChatGPT Browse with Bing은 NYT 저작물로부터 생성된 합성 검색결과 형태로 NYT 저작물의 복제물 및 2차적저작물을 생성한다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 침해: NYT 웹사이트들로부터 NYT 저작물을 스크래핑하고 제3자의 데이터셋으로부터 NYT 저작물을 재복제함으로써, 수백 만 건의 NYT 저작물의 복제물을 포함하는 학습 데이터셋을 구축하여 저작권을 침해하였다.

21) MS가 Bing 검색엔진에 통합해 제공했던 대화형 AI 검색 인터페이스였는데, 현재는 브랜드가 Copilot으로 통합·개편되었음. 곧 이용자가 질문을 입력하면 Bing 인덱스는 관련 웹페이지를 검색하고, 검색된 콘텐츠를 직접 그대로 나타내거나 표현을 바꾸어 서술(패러프레이즈)하여 답변을 제공.

22) ChatGPT가 실시간 웹 검색을 할 수 있도록 Bing 검색엔진을 호출해 답변을 생성하는 기능. 곧 이용자가 질문을 입력하면, ChatGPT가 Bing 검색을 호출하고, Bing 검색 인덱스에서 웹 콘텐츠를 검색·가져오는데, 콘텐츠를 그대로 나타내거나 표현을 바꾸어 서술(패러프레이즈)하여 답변을 제공.

- ② 대위 침해: MS는 ①OpenAI에 의한 저작권 침해를 통제·지휘(direct)하고 이로부터 이익을 얻었고, ②학습 데이터셋, GPT 모델, OpenAI의 ChatGPT 제공 서비스를 저장·처리·복제하는 데 사용되는 슈퍼컴퓨팅 플랫폼을 통제·지휘하고, OpenAI 피고들에 의한 저작권 침해로부터 이익을 취하였다.
- ③ 기여 침해: MS는 ①OpenAI의 직접 저작권 침해에 실질적으로 기여하고 이를 직접적으로 지원하였고, OpenAI의 침해를 알고 있었거나 최소한 알만한 이유가 있었고, ②이용자의 저작권 직접 침해에 실질적으로 기여하고 이를 직접 지원하였으며, 이용자에 의한 직접 침해를 알고 있었거나 알만한 이유가 있었다.
- ④ 저작권 관리정보의 제거: NYT는 각 저작물마다 저작권 표시, 저작물 제호 및 기타 저작물 식별 정보, 이용 조건, 저작권 관리정보에 대한 식별번호나 기호 등 CMI(Copyright Management Information, 이하 CMI)를 포함하고 있는데, (a)피고들은 스크래핑된 NYT의 저작물과 제3자 데이터셋에서 복제한 NYT의 복제물에서 CMI를 제거하였다. §1202(b)(1) 위반). MS와 OpenAI는 합성 검색 결과를 제공하는 과정에서도 NYT의 CMI를 제거하였다. 또한 (b)CMI가 없이 NYT의 복제물 및 2차적저작물을 AI 결과물로 제공하였다(§1202(b)(3) 위반).

13. Authors Guild, et al. v. OpenAI Inc.

(1) 뉴욕 남부 연방지방법원(1:23-cv-08292, 2023.9.19.)

원고는 작가 조합(Authors Guild) 및 존 그리삼(John Grishman)과 조지 R.R. 마틴(George R.R. Martin)을 포함한 수십 명의 작가들이고, 피고는 다수의 OpenAI 법인으로 피고가 소설 작가 및 저작권 소유자의 권리를 침해했다며 제기된 집단소송이다.

(2) 사실관계(원고 주장)

- ① 원고들은 작가 조합(Authors Guild)²³⁾과 픽션 작품을 창작하는 작가들이다.
- ② 피고 OpenAI는 GPT-3, GPT-3.5, GPT-4 등 자신의 LLM을 학습시키기 위하여 원고들의 서적을 복제하였다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 직접 침해: OpenAI는 LLM 및 ChatGPT를 학습시키기 위하여 원고들과 집단 구성원들의 저작물을 복제함으로써 저작권을 침해하였다.
- ② 저작권 대위 침해: OpenAI Inc.는 관련 회사인 OpenAI GP LLC 등을 완전히 지배하고,²⁴⁾ 이들의 직접 침해 행위를 통제할 권리와 능력을 보유하고 있었고, OpenAI LP의 침해 행위로부터 발생한 수익과 투자로부터 이익을 얻고 있으므로, 직접 침해에 대하여 직접적인 금전적 재정적 이해관계를 가지고 있다.

23) 미국에서 가장 오래되고 규모가 큰 작가 전문 단체로서 표현의 자유와 저작권 보호를 옹호하고, 전문 작가들의 이익을 대변하는 전문 단체(association)임.

24) 이 케이스와 OpenAI를 상대로 하는 케이스들에서는 피고로 OpenAI 외에도 OpenAI GP LLC, OpenAI LP, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation LLC, OpenAI Holdings LLC 등이 등장하는데, 이들 기업들은 OpenAI를 최상층으로 하여 이루어진 계열회사라고 할 수 있음.

14. The Intercept Media, Inc. v. OpenAI, Inc.

(1) 뉴욕 남부 연방지방법원(1:24-cv-01515, 2024.2.28.)

원고는 미국의 비영리 뉴스 매체인 The Intercept를 운영하는 업체이고, 피고는 다수의 OpenAI 법인으로 양질의 뉴스 콘텐츠를 OpenAI의 모델 학습에 무단으로 사용했다며 제기된 소송이다.

(2) 사실관계(원고 주장)

- ① 원고 The Intercept Media, Inc.는 언론 기관으로서, 원고의 저널리즘 저작물은 인터넷에 게시되어 있고, 저작자, 제호, 저작권, 이용 조건에 관한 정보(CMI)와 함께 공중에게 전달된다.
- ② ChatGPT는 이러한 CMI 없이 저작물을 모방한 자료를 이용자에게 제공하고, 수천 건의 원고 저작물들이 CMI 없이 피고들의 학습 데이터셋에 포함되었다. MS는 ChatGPT 및 Bing Copilot 학습 데이터셋에서 원고의 저작물로부터 CMI를 고의로 제거하였고, ChatGPT 개발 노력의 일환으로, CMI가 없는 원고 저작물의 복제물을 OpenAI와 공유하였다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 관리정보 제거 금지에 관한 규정 위반(OpenAI): OpenAI는 CMI가 제거된 원고 저작물의 복제물을 제작하여 학습 데이터셋에 포함시켰고, CMI 제거 행위가 이용자의 저작권 침해를 가능하게 하거나, 침해를 은폐할 것임을 알거나 알만한 이유가 있었다(§1202(b)(1) 위반).
- ② 저작권 관리정보가 제거·변경된 저작물의 배포 등을 금지하는 규정 위반(OpenAI): OpenAI는 ChatGPT 개발과 관련하여 CMI가 제거된 원고 저작물의 복제물을 피고 MS와 공유하였다(§1202(b)(3) 위반).
- ③ 저작권 관리정보 제거 금지에 관한 규정 위반(MS): MS는 CMI가 제거된 원고의 저작물을 작성하여 ChatGPT 및 Bing AI 제품의 학습 데이터셋에 포함시켰다(§1202(b)(1) 위반).
- ④ 저작권 관리정보가 제거·변경된 저작물의 배포 등을 금지하는 규정 위반(MS): MS는 ChatGPT 개발과 관련하여 CMI가 제거된 원고 저작물의 복제물을 OpenAI와 공유하였다(§1202(b)(3) 위반).

15. The Center for Investigative Reporting, Inc. v. OpenAI et al. and Microsoft

(1) 뉴욕 남부 연방지방법원(1:24-cv-04872, 2024.6.27.)

원고는 비영리 뉴스 기관인 탐사보도센터(The Center for Investigative Reporting)이고, 피고는 OpenAI 관련 법인 및 마이크로소프트사이다.

(2) 사실관계(원고 주장)

- ① 원고 The Center for Investigative Reporting, Inc.(CIR)는 미국에서 가장 오래된 비영리 뉴스 매거진 및 웹사이트이다.

- ② ChatGPT 버전들은 최소한 WebText, WebText2, Common Crawl에서 파생된 데이터셋을 사용하여 학습되었는데, WebText와 WebText2는 OpenAI가 생성한 것이다. OpenAI는 WebText를 개발하는 과정에서 Dragnet과 Newspaper라는 알고리즘 세트를 사용하여 웹사이트에서 텍스트를 추출하였는데, 이 알고리즘은 원고 웹사이트의 푸터(footer)²⁵에 포함된 저작자 정보, 저작물 명칭(제호), 저작권 고지 및 이용약관 정보를 제거하는 것이다.
- ③ ChatGPT와 Copilot은 이용자에게 저작물을 그대로(verbatim) 또는 거의 그대로 역류시키고(regurgitation), 그러면서도 CMI를 제공하지 않았다. 또한 ChatGPT는 학습데이터를 기억(memorize)하고 콘텐츠를 역류시킨다.
- ④ ChatGPT와 Copilot은 이용자의 프롬프트에 응하여 원고의 저작물에 대하여 매우 상세한 축약본(abridgements)을 제공한다.

(3) 침해(위반) 주장 및 청구 사항

- ① 저작권 직접 침해: (a)인터넷에서 저작물을 다운로드하고, 컴퓨터 메모리에 인코딩(encoding)하고, (b)이용자들의 프롬프트에 응하여 등록 저작물을 그대로 또는 거의 그대로 역류시키고, (c)프롬프트에 응답하여 저작물을 축약함으로써, 저작권을 침해하였다.
- ② 저작권 간접 침해(기여 침해): 이용자들의 직접 침해에 대하여 실질적으로 기여하고 직접적으로 조력하였고, 이용자들의 직접 침해를 알고 있었거나 알만한 이유가 있었다.
- ③ 저작권 관리정보(CMI) 규정 위반: (a)CMI가 제거된 원고 저작물을 학습 데이터셋에 포함시키고, 이용자에게 의한 저작권 침해를 가능하게 하거나 용이하게 하거나 은폐할 것을 알만한 이유가 있었다(§1202(b)(1) 위반). (b)OpenAI와 MS는 CMI가 제거된 원고 저작물의 복제물을 서로 공유하였다(§1202(b)(3) 위반).

25) 웹페이지 하단 영역을 의미하는데, 흔히 저작권 고지(© 2024 Michael Fox), 이용약관, 라이선스 요건, 권리자 정보 등 CMI가 표시됨.

참고 자료

- <https://copyrightalliance.org/wp-content/uploads/2025/07/DENIAL-v.-OPENAI-COMPLAINT-JUNE-30-2025.pdf>
- https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.598.0_1.pdf
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.644782/gov.uscourts.nysd.644782.1.0.pdf>
- https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0_4.pdf
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.616533/gov.uscourts.nysd.616533.178.0.pdf>
- https://www.musicbusinessworldwide.com/files/2025/06/JUSTICE-et-al-v.-UNCHARTED-LABS-INC.-d_b_a-Udio.com_-Entry-1.pdf
- <https://chatgptiseatingtheworld.com/wp-content/uploads/2025/06/Justice-v.-Suno-June-16-2025.pdf>
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.606655/gov.uscourts.nysd.606655.456.0.pdf>
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.606655/gov.uscourts.nysd.606655.456.0.pdf>
- <https://www.courtlistener.com/docket/67538258/tremblay-v-openai-inc/>
- https://storage.courtlistener.com/recap/gov.uscourts.nysd.612697/gov.uscourts.nysd.612697.1.0_1.pdf
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.606655/gov.uscourts.nysd.606655.456.0.pdf>
- https://storage.courtlistener.com/recap/gov.uscourts.nysd.616536/gov.uscourts.nysd.616536.1.0_1.pdf
- <https://storage.courtlistener.com/recap/gov.uscourts.nysd.623941/gov.uscourts.nysd.623941.1.0.pdf>